95-845: AAMLP Project

Heinz 95-845: Project Report

Mridul Gangwar MGANGWAR@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University, Pittsburgh, PA, United States

Lauren Rost LROST@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University, Pittsburgh, PA, United States

Nikita Setia NIKITAS@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University, Pittsburgh, PA, United States

Abstract

Opioid overdose deaths spiked in 2017 and continue to be a concern for the US healthcare system. Major efforts have been dedicated to understand the demographic impacted and how the healthcare system can improve outcomes for individuals at risk. This paper explores the application of machine learning and statistical analyses to identify which could best predict and prevent overdose deaths using county-provided demographic, program activity and opiate prescription fills data for Medicaid beneficiaries from 2009-2017. The models with superior performance, Multivariate Logistic Regression and Random Forest, yield an AUROC of approximately 0.83, correctly identifying 57% of individuals who overdose. We found that the number of Tramadols prescribed, the number of drug and alcohol services received, whether an individual is White, and the number of lower court drug-related criminal cases were the most informative variables towards predicting opioid overdose deaths. We also note survival probabilities varying by gender, age and race. This paper showcases how county-provided data can be used to predict and prevent overdoses, translatable to other counties, identifies the variables indicative of risk, and highlights modeling techniques that can be tuned to improve predictive performance, thereby moving the field towards overdose prevention.

1. Introduction

Opioid abuse has been an emerging public health issue, and was declared a public health emergency in 2017[1]. In the past two decades, national drug overdose deaths have increased from 16,849 in 1999 to 70,237 in 2017[2]. This trend is especially affected by the onset of the opioid epidemic in the late 1990s[3]. Approximately 67% of the overdose deaths in 2017 were due to the involvement of any opioids (including illegal drugs like heroin and fetanyl) and 24% are specifically due to legally prescribed opioids, like oxycodone, hydrocodone, codeine, morphine, etc.[4][5]. It has become critical to better understand the risk factors leading to overdoses and to determine the best way to prevent overdose deaths.

Crosier et al. predicted overdose frequency using random forests in order to uncover important features related to the frequency and development of overdose events[6]. Lobo et al. identified subgroups of Pennsylvania patients at greater risk for opioid abuse in a k-means clustering algorithm[7]. Our work aims to contribute to this existing body of work by identifying features that are most indicative of risk, which is crucial to preventing addiction and overdoses. There exists an Opioid Risk Tool, developed in 2005, to flag patients at risk for opioid abuse and overdose[8]. However, due to the subjective nature of this tool and the spike of deaths in 2017, machine learning has been sought to provide a more objective and quantitative approach to estimate risk of opioid abuse and overdose. A super learning approach was developed by Acion et al. to predict the successful treatment of patients with substance use disorders[9]. Artificial intelligence was also applied in the sphere of predicting opioid abuse by Haller et al., who implemented natural language processing on

electronic health record data to assess risk and predict opioid abuse[10].

This paper strives to expand upon previous contributions to the field by identifying the machine learning algorithms that yield the highest predictive performance of overdose events. Specifically, those models that correctly identify the greatest number of at-risk individuals. This project uses datasets containing information concerning demographics, county-provided program usage, and opioid prescriptions to predict overdose deaths, opioid and non-opioid. Specifically, we analyze the data of Medicaid beneficiaries from 2009 to 2017 provided by the Allegheny County Department of Human Services (DHS) to predict the risk of death due to an non-opioid or opioid-related overdose. Here, we apply machine learning to uncover features and methods that successfully predict overdose death, and thereby enhance the space of addiction and overdose death prevention.

2. Methods

2.1 Original Data Description

We accessed 3 datasets from Allegheny County DHS: demographic, program activity, and opiate prescription fills. They contain data for 120,650 individuals who have utilized DHS services between 2009 and 2017. The demographic dataset (summarized in appendix figure 6) contained variables for person ID, race, and gender. Of note, there was missingness in terms of race data for 19,531 individuals and gender information for 210 individuals. Additionally, there were demographics that were not heavily represented in the rest of the dataset: one transgender female and 19 Native Hawaiian / Pacific Islanders.

The program dataset (summarized in appendix figure 7) contains 2,402,479 rows where each row denotes the activity, or activities, for an individual at any time since they entered the system. The program dataset contains variables for person ID, year and month of activity, overdose details (if any), and the DHS program-related relevant activity information. Program-related activity information included whether there was documentation for Child, Youth and Family (CYF) services used as child or parent (binary); the number of criminal court cases (drug or not) filed; mental health, drug and alcohol abuse, or prescription services (binary); and whether the individual was jailed (binary). If an individual experienced an overdose event, there were additional data fields for overdose date and a binary value for whether the overdose was opioid-related. The only missingness in the program activity dataset was in the overdose-related fields for individuals who did not overdose.

The opiate prescription fills dataset (summarized in appendix figure 8) contains 1,161,650 rows, where each row represents a prescription fill for an individual. The prescription information variables are the claim number (unique for each row), person ID, age at prescription, dispensed quantity, days supply, fill date, and information specific to the drug (drug strength, name variations, package description, and dosage form). We did not include generic tier description and claim rank of prescriptions in our analyses for they are not informative. There were missing, extreme or incomprehensible values in the prescription dataset, such as an age of -7990, dispensed quantity of 17936 and days supply of 907. In addition, the prescription dataset contained multiple versions of drug names for the same drug. There were two columns pertaining to the drug dosage form: a condensed version and a descriptive version.

2.2 Data Cleaning and Feature Extraction

2.2.1 Demographic and Program Datasets

The missing race and gender information in the demographic dataset is likely missing not at random (MNAR) as it may be directly related to the unreported value itself. We dealt with this missingness

through assigning missing values their own category, "No Data". Furthermore, with respect to the race column, given that the "Native Hawaiian / Pacific Islander" demographic was underrepresented with only 19 individuals, we merged these individuals with the "No Data" race category to create a variable called "No Data and Other". This was replicated for the gender variable where we merged the "Transgendered male to female" and the "No Data" categories into "No Data and Other." With respect to the program activity dataset, we cleaned the outcome variable by converting the 0, 1 and NA to "Non-Opiate Overdose", "Opiate Overdose" and "No Overdose".

2.2.2 Opiate Prescription Fills Dataset

The age at which an individual got their first prescription was extracted. If the minimum age was less than 0, the maximum age was used. If both values were less than 0, then they were considered missing. There were 26 individuals with missing age values; these were replaced with the median age of individuals in the entire dataset. Drug strength values of oral solutions were updated to 5-325/5ML and of Naloxone were updated to 50MG-0.5MG. Days supply and dispensed quantity extreme values were censored. The top extreme values were replaced with the value at the 0.5% percentile and bottom extreme with the value at the 95.5% percentile. These percentile values were determined for each generic name and dosage form combination. Opioid strength values were pulled from the drug's label name, normalized to be per 1 ML, MCG or MG (if applicable). Missing values were replaced with those from provided drug strength column. Drug name (generic) and dosage forms were stripped to common terminology. The above-mentioned opioid strength values were converted into Morphine Milligram Equivalents (MME) using Opioid Morphine Equivalent Conversion Factors[11]. The formula used was: $opioid\,strength * (dispensed\,quantity/days\,supply) * conversion\,factor$.

2.2.3 Final Dataset

There were 30 meaningful features that were extracted at the person ID level. Demographic characteristic features included race, gender and age. DHS program usage features entailed cohort (the enrollment year in DHS services), usage of CYF program as child ($total\_cyfchild$) and as parent ($total\_cyfparent$), use of mental health, drug and alcohol abuse and prescription services ($total\_mh$, $total\_da$, $total\_rx$), number of months spent in jail ($total\_acj$), and number of criminal cases ($total\_cr\_cases$) and drug-related cases filed in court ($total\_cr\_drug\_cases$).

Prescription activity features extracted included the number of prescriptions (num_presc), name of the most prescribed drug ($most\_presc\_drug$) and most prescribed dosage form ($most\_dose\_form$), count of top three drugs (oxy_count, $tram\_count$, $hydrobit\_count$), count of top three dose forms ($pill\_count$, $patch\_count$, $liquid\_count$), average, median and mode MME (avg_mme, $median\_mme$, $mode\_mme$), as well as average days supply and dispensed quantity (avg_supply and $avg\_dispensed$). Finally, overdose information (od_type, od_month, od_year, od_date) was added.

2.2.4 Feature Choices

Some additional feature choices were made to enhance the final dataset and prepare for the prediction tasks. Features such as $most\_presc\_drug$ and $most\_dose\_form$ had categories that were comprised of less than 0.2% of the values. To prevent errors during the prediction task, these low occurrence categories were combined to create a separate Other category. We plotted the distribution of the average, median, and mode of morphine milligram equivalents (MME) variables to determine which mode of central tendency for MME was normally distributed. This led us to retain only median MME in the final dataset to avoid multicollinearity.

We examined the outcome variable under a binary value of 1 for an "opiate overdose" or "non-opiate overdose" and 0 for there being "no overdose". We include non-opiate overdose events in our

outcome variable because these individuals were receiving opioid prescriptions and we are willing to increase the number of potential false positives to flag as many potentially at-risk individuals as possible. Furthermore, we argue that individuals who overdose (opiate or non-opiate) share similar characteristics and so the inclusion of both types of overdose in our outcome variable is informative in the prediction task. Columns pertaining to the date of overdose were removed to avoid leakage in the machine learning models, as they are proxies for the target variable. All categorical variables were converted to dummy variables and all variables were log transformed.

2.3 Oversampling, Machine Learning and Statistical Models, and Evaluation Metrics

Since there is a severe under-representation of outcome variable events (1,222 out of 120,650 individuals), the first model (multivariate logistic regression) was unable to capture a single case of overdose. As such, this paper utilized an oversampling method, Synthetic Minority Oversampling Technique (SMOTE) for the outcome variable. To do so, we divided the whole dataset into train and test with a 50-50% split. We then ran SMOTE on the train set to generate synthetic data to train our models and then calculated performance on the original test set. Furthermore, to increase the performance of our machine learning models, we trained the data only on the top 20 critical features identified using the random forest algorithm.

Six machine learning models were run and their performance in predicting overdose rates compared. These models include: multivariate logistic regression, ridge regression, random forest, AdaBoost, gradient boosting, and neural networks. We used 5-fold cross-validation for multivariate logistic regression, and 10-fold cross-validation for ridge regression, random forest, and gradient boosting. We also implemented lasso regularization. Note that for this use case in particular, it is more critical to identify the individuals who may overdose (true positives) than to eliminate individuals not at risk from consideration (true negatives). Furthermore, given the limited resources available to tackle such a pressing issue, it is valuable to minimize (to the extent possible) the number of individuals incorrectly identified to be at risk (false positives). Therefore, we evaluated and compared machine learning models through the area under the curve (AUC) of the Receiver Operating Characteristic (ROC) curves, the number of true positives, the number of false positives, sensitivity, specificity, and precision. The important or significant features used in these models will also be an additional point of comparison, specifically for the multivariate logistic regression and random forest models.

This paper also includes a survival analysis, conducted using Kaplan-Meier estimation and Cox Proportional Hazards models. For the survival analysis, the time variable was the months in system and the event variable was the status of the individual at their last known month in the system. The maximum month in the system is 108 months (December 2017). The event was 1 for overdose death and 0 otherwise. The Kaplan-Meier estimation was evaluated visually (with the survival curve) along with the model output. Both normal and regularized versions of the Cox Model were produced. The coefficients, exponent coefficients and their corresponding p-values were used to understand and evaluate both models. Cross-validation was used to determine the best lambda to use for the regularized Cox Model.

3. Results

Table 1 provides an overview of the 120,650 individuals in the Allegheny County DHS cohort between 2009 and 2017. No individuals were excluded from the cohort. We addressed all missingness to the extent possible.

Table 1. Demographics of Allegheny County DHS Cohort (2009-2017).

Characteristic	N	Percentage (%)
Race		
White	56,754	47.04
Black/African-American	40,573	33.63
Biracial/Multiracial	2,118	1.75
Asian	1,216	1.00
American Indian/Alaskan Native	439	0.36
No Data and Other	19,550	16.20
Total	**120,650**	**100**
Gender		
Female	73,582	60.99
Male	46,857	38.84
No Data and Other	211	0.17
Total	**120,650**	**100**
Age(years)		
0-19	27,128	22.48
20-39	51,883	43.00
40-59	35,941	29.79
60 & Over	5,698	4.72
Total	**120,650**	**100**

3.1 Evaluation of Machine Learning Models

The ROC curves of all 6 machine learning models run in this paper are showcased in figure 1 below.

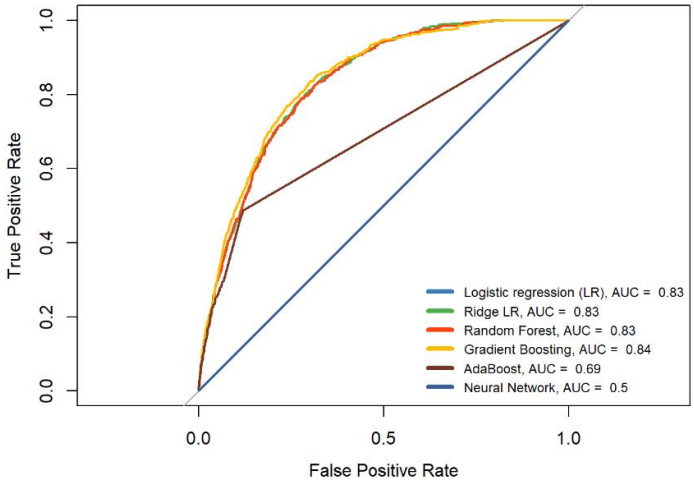

Figure 1. Receiver Operating Characteristic (ROC) curves with corresponding AUCs for the 6 ML models

These curves, along with their reported AUCs, indicate that Gradient Boosting performs the best with logistic regression, ridge regression, random forest close behind. AdaBoost does not perform well and neural networks perform only as good as chance. As such, AdaBoost and neural networks can be dropped from consideration for the time being. However, it would be premature to select gradient boosting as the best model. As stated earlier, it is vital to identify the individuals who are at risk of overdose (true positives) while keeping the number of false identifications to a minimum. As such, while gradient boosting may appear to perform the best, whether it is truly the best model in this case depends on its performance on additional evaluation metrics.

Table 2 summarizes the performance of all six machine learning models in predicting overdose deaths. Note that there are 633 overdoses in the test dataset. As seen in this table, despite having the highest AUC, gradient boosting correctly identifies only 22% of overdose deaths (few relative to

Table 2. Evaluation Metrics of the Machine Learning Models

Model	True Positives	False Positives	Sensitivity (%)	Specificity (%)	Precision (%)	AUC
Multivariate Logistic Regression	362	8654	57	86	4	0.83
Ridge Regression	163	2422	26	96	6	0.83
Random Forest	362	8654	57	86	4	0.83
Gradient Boosting	138	1982	22	97	7	0.84
Ada Boost	147	2389	23	96	6	0.69
Neural Net	633	59692	100	0	0	0.5

the other high performing models). Neural networks identifies all overdose deaths simply because it classifies every individual at risk of overdose, so it can be excluded from consideration. Of the remaining four models, multivariate logistic regression and random forest perform the best, correctly identifying 57% of the at-risk individuals.

3.2 Understanding the Survival Analysis

Survival analysis using the Kaplan-Meier estimation yields a survival curve seen in figure 2 in the appendix. The plot shows us that the probability of survival gradually declines from 100% at the beginning to 96.6% at the final month in the system (108 month mark). Individuals have dropped out i.e. stopped utilizing DHS services at various points throughout. The earliest time at which we start to see a decline in survival probability is 8 months.

Next, we explored if and how this probability of survival changes among various gender and race subgroups. Figure 3 shows us that females have a higher probability of survival than males in this dataset. Furthermore, individuals who are in the "No Data and Other" category have a relatively lower probability of survival. Females have a 97.5% probability of surviving, males have a 95.3% probability of surviving, and individuals who have either not reported their gender or are transgender female have a 72.7% probability of surviving an overdose event by the time they conclude their time in the system.

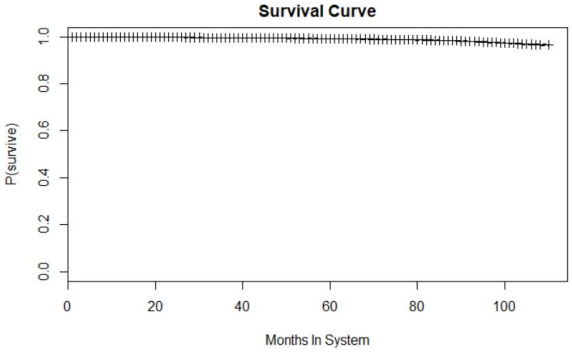

Figure 2. Kaplan Meier survival curve (all).

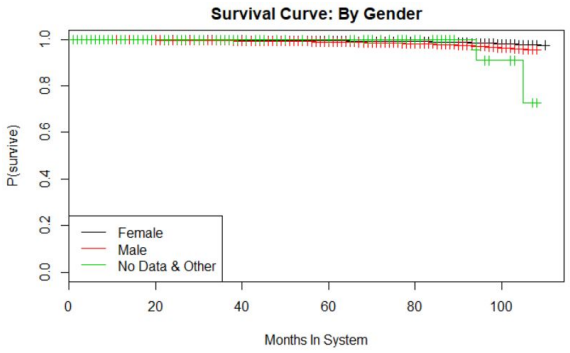

Figure 3. Kaplan Meier curve by gender.

Table 3 below displays the survival probabilities at the conclusion of time in the system by race. Since no American Indian or Native Alaskan individuals experience an overdose event, no survival probability was calculated for them. There appear to be subtle differences in survival probabilities among the remaining racial categories. For example, we see that White individuals have the lowest relative survival probability (95%) and Asian individuals have the highest survival probability (99.6%). None of the probabilities, however, are lower than 95% and the differences are not as noticeable as those seen by gender alone.

Next, we use a Cox Proportional Hazards model to analyze the survival probability of individuals by including age, gender and race as predictor variables. Figure 4 below shows the output of the Cox

Table 3. Probability of survival by Race

Race	Probability of Survival (%)
American Indian or Native Alaskan	Not Applicable
Asian	99.6
Biracial or Multiracial	96.6
Black or African American	98.6
No Data or Other	97.4
White	95.0

model which indicates that age, being male, being in the "No Data and Other" gender category, and being White are statistically significant predictors of survival. The base variables are being female and being Asian. Given that there are no overdose events for the American Indian or Native Alaskan individuals, their data was merged into the "No Data and Other" race category so as to not skew the results.

The results show that, **holding all other things constant**: a one year increase in age is associated with an increase in hazard by a factor of 1.028 or 2.8%; compared to being female, being male increases hazard by a factor of 1.87 or 87% and being an individual with no reported gender or other increases hazard by a factor of 7.9 or 690%; and compared to being Asian, being White increases hazard by a factor of 8.2 or 720%. These findings are aligned with our findings using the Kaplan Meier estimation.

We also ran a regularized Cox Proportional Hazard model and used cross-validation to determine the best (minimum) lambda value. Please refer to table 4 in the appendix for the resulting coefficients and associated change in hazard (%). These findings are also consistent with the Kaplan Meier survival probabilities.

```
                              coef exp(coef) se(coef)     z      p
age                        0.02736   1.02774  0.00208 13.19 <2e-16
genderMale                 0.62619   1.87047  0.05815 10.77 <2e-16
genderNo Data and Other    2.06827   7.91115  0.71324  2.90 0.0037
raceBiracial/Multiracial   1.55887   4.75345  1.03539  1.51 0.1322
raceBlack/African-American 0.81961   2.26961  1.00269  0.82 0.4137
raceNo Data and Other      0.91127   2.48747  1.01580  0.90 0.3697
raceWhite                  2.10949   8.24405  1.00066  2.11 0.0350
```

Figure 4. Cox Model Output with Age, Race and Gender as Predictors

3.3 Variable Importance

Figure 5 below showcases the top 15 variables from the Random Forest model. The variables most important in predicting overdose was the total number of Tramadol prescribed ($tram\_count$), the total number of drug and alcohol services received ($total\_da$), whether an individual is white ($race\_White$), and the total number of lower court drug-related criminal cases ($total\_cr\_drug\_cases$). These four variables were also statistically significant in the multivariate logistic regression. Most of the remaining variables were statistically significant except the average number of dispensed drugs ($avg\_dispensed$), whether an individual was male ($gender\_Male$), and the total number of mental health services received ($total\_mh$).

4. Discussion

Due to the under-representation of overdose events in our dataset, in spite of oversampling, even the best machine learning models yielded a relatively low sensitivity (57%) and fairly high number of

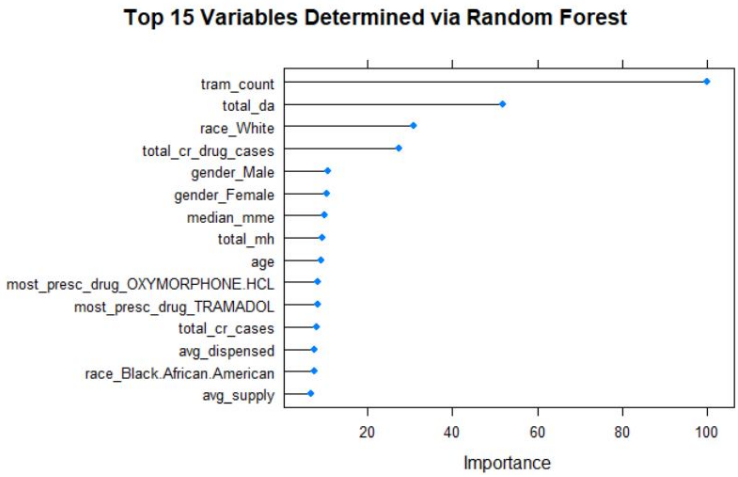

Figure 5. The top 15 variables of importance elucidated from the random forest model.

false positives. While one could argue that it is beneficial to be extra cautious by overestimating the number of at-risk individuals as it would go a long way towards preventing addiction and overdose deaths. However, the high false positives may overburden counties' substance abuse prevention and treatment efforts. One way to address this could be gathering additional data, perhaps by pooling data from other counties or for additional years, or both. Another way to enhance the existing algorithms (especially boosting) would be to tune the hyperparameters. Moreover, the neural network had very poor performance because we utilized four dense layers and ran the algorithm for 30 epochs. While a deep neural network was computationally expensive for us, we would expect it to yield a higher performance than the one seen in this paper.

While our analysis showcased gender- and race-related differences in survival, we cannot be certain that it is not merely a reflection of the Allegheny County population proportions or if it is reflective of some association between these factors and drug abuse. Our work may also be limited by the fact that it may not be generalizable to other counties, especially if they do not collect similar data in the same way. Furthermore, certain features were created specifically for this dataset, such as the count of the top three drugs (one of which was very predictive of overdose events) and dosage forms. Our analysis is also limited by the data collected by the county's DHS; we did not have access to potentially relevant variables, such as socioeconomic status, medical history, diagnoses, and insurance claims. Additionally, some feature creation could benefit from medical expertise, such as MME.

5. Conclusion

This paper applies the analytical pipeline on Allegheny County DHS demographic, prescription and program activity data for 120,650 individuals from 2009 to 2017 to predict the rate of opiate and non-opiate overdose. The best models, logistic regression and random forest, yielded an AUC of 0.83 and correctly identified 57% of the at-risk population. We believe this analytical pipeline is a foundation that can be further improved using a bigger, enriched dataset that has a relatively more balanced outcome variable, better tuned models, and additional computational power, thereby moving the field one step closer towards preventing addiction and overdose deaths.

Appendix

Code is available at https://github.com/nikitascmu/S19-aamlp-project-group1.

```
   PERSON_ID                                    RACE                                  GENDER
 Min.   :     1   American Indian/Alaskan Native  :  439   Female                       :73582
 1st Qu.: 30163   Asian                           : 1216   Male                         :46857
 Median : 60326   Biracial/Multiracial            : 2118   No Data                      :  210
 Mean   : 60326   Black/African-American          :40573   Transgendered male to female:    1
 3rd Qu.: 90488   Native Hawaiian/Pacific Islander:   19
 Max.   :120650   No Data                         :19531
                  White                           :56754
```

Figure 6. Summary of the original demographic dataset.

```
  PERSON_ID           YEAR          MONTH          CYFCHILD         CYFPARENT       MDJS_CR_CASES      MDJS_CR_DRUG_CASES
 Min.   :     1   Min.   :2009   Min.   : 1.00   Min.   :0.00000   Min.   :0.00000   Min.   : 0.00000   Min.   :0.00000
 1st Qu.: 30364   1st Qu.:2011   1st Qu.: 3.00   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.: 0.00000   1st Qu.:0.00000
 Median : 60557   Median :2013   Median : 6.00   Median :0.00000   Median :0.00000   Median : 0.00000   Median :0.00000
 Mean   : 60425   Mean   :2013   Mean   : 6.48   Mean   :0.02653   Mean   :0.07775   Mean   : 0.04274   Mean   :0.01301
 3rd Qu.: 90391   3rd Qu.:2015   3rd Qu.: 9.00   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.: 0.00000   3rd Qu.:0.00000
 Max.   :120650   Max.   :2017   Max.   :12.00   Max.   :1.00000   Max.   :1.00000   Max.   :18.00000   Max.   :6.00000

       MH                DA                RX                ACJ          OVERDOSE_DATE     OPIATE_OVERDOSE
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :200901   Min.   :0.0
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:201411   1st Qu.:1.0
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.00000   Median :201607   Median :1.0
 Mean   :0.4921   Mean   :0.1509   Mean   :0.3453   Mean   :0.07592   Mean   :201539   Mean   :0.9
 3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:201703   3rd Qu.:1.0
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.00000   Max.   :201802   Max.   :1.0
                                                                      NA's   :2362739   NA's   :2362739
```

Figure 7. Summary of the original program activity dataset.

```
      CLAIM_NR          PERSON_ID           AGE          DISPENSED_QTY       DRUG_STRENGTH     DAYS_SUPPLY
 000000064742:      1   Min.   :     1   Min.   :-7990.00   Min.   :    0.00   50 MG     :123829   Min.   :  0.00
 000000064849:      1   1st Qu.: 29949   1st Qu.:   32.00   1st Qu.:   18.00   5MG-325MG : 89846   1st Qu.:  4.00
 000000065039:      1   Median : 59718   Median :   45.00   Median :   40.00   5 MG      : 87096   Median : 11.00
 000000066034:      1   Mean   : 59960   Mean   :   42.69   Mean   :   58.73   5 MG-325MG: 71169   Mean   : 15.57
 000000070555:      1   3rd Qu.: 89653   3rd Qu.:   53.00   3rd Qu.:   90.00   7.5-750MG : 65156   3rd Qu.: 30.00
 000000115549:      1   Max.   :120650   Max.   :  354.00   Max.   :17936.00   5MG-325M  : 56623   Max.   :907.00
 (Other)     :1161644                    NA's   :126                           (Other)   :667931
     FILL_DATE                                LABEL_NAME                BRAND_NAME                      GENERIC_NAME
 2013-02-01:    868   TRAMADOL HCL 50 MG TABLET     :148444   HYDROCODONE-ACETAMI:495685   HYDROCODONE BITARTRATE:498280
 2013-07-01:    863   OXYCODONE-ACETAMINOPHEN 5-325 :147962   OXYCODONE-ACETAMINO:190069   OXYCODONE HCL/ACETAMIN:218559
 2014-08-01:    859   HYDROCODON-ACETAMINOPHEN 5-500:132225   OXYCODONE HCL      :176250   OXYCODONE HCL          :184272
 2012-06-01:    855   HYDROCODON-ACETAMINOPH 7.5-750:120385   TRAMADOL HCL       :148444   TRAMADOL HCL           :148635
 2013-08-01:    841   HYDROCODON-ACETAMINOPHEN 5-325:112172   MORPHINE SULFATE ER: 26912   MORPHINE SULFATE       : 31889
 2013-04-01:    837   OXYCODONE HCL 5 MG TABLET     : 93667   FENTANYL           : 21559   OXYMORPHONE HCL        : 28456
 (Other)   :1156527   (Other)                       :406795   (Other)            :102731   (Other)                : 51559
    PACKAGE_DESC     DOSAGE_FORM_CD                                DOSAGE_FORM_DESC   GENERIC_TIER_CLASS_DESC   CLAIM_RANK
 BOTTLE    :1138114   TA     :1059465   TABLET                              :1059465   ANALGESICS:1161650      Min.   :1
 BOX       :  22198   TS     :  26805   TABLET, EXTENDED RELEASE            :  26805                           1st Qu.:1
 BLIST PACK:    843   PR     :  21677   PATCH, TRANSDERMAL 72 HOURS         :  21677                           Median :1
 VIAL      :    213   TM     :  19707   TABLET, EXTENDED RELEASE 12 HR      :  19707                           Mean   :1
 SQUEEZ BTL:    198   SJ     :  14668   SOLUTION, ORAL                      :  14668                           3rd Qu.:1
 AMPUL     :     45   FT     :  14305   TABLET,ORAL ONLY,EXTENDED RELEASE 12:  14305                           Max.   :1
 (Other)   :     39   (Other):   5023   (Other)                             :   5023
```

Figure 8. Summary of the original opiate prescription fills dataset.

Table 4. Regularized Cox proportional Coefficients

Race	coefficient	exp(coefficient)	Change in hazard as %
age	0.027	1.027	2.7
gender_female	-0.617	0.540	-46
gender_male	.	.	.
gender_No Data and Other	1.285	3.615	261.5
race_white	1.151	3.161	216.1
race_Black/African-American	-0.119	0.888	-12.2
race_Biracial/Multiracial	0.543	1.721	72.1
race_No Data and Other	.	.	.
race_Asian	-0.459	0.632	-36.8
race_American Indian/Alaskan Native	-1.102	0.332	-66.8

Endnotes

[1]Health and Services (2019)
[2]National Institute On Drug Abuse (2019a)
[3]National Institute On Drug Abuse (2019b)
[4]National Institute On Drug Abuse (2019a)
[5]on Drug Abuse (2019)
[6]Sage Crosier (2017)
[7]Lobo (2017)
[8]Webster and Webster (2005)
[9]Acion (2017)
[10]Haller (2017)
[11]for Medicare and (CMS)

References

Kelmansky D. Laan M. V. Sahker E. Jones D. Arndt S Acion, L. Use of a machine learning framework to predict substance use disorder treatment success, April 2017. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0175383.

Centers for Medicare and Medicaid Services (CMS). Opioid oral morphine milligram equivalent (mme) conversion factors, 2017. URL https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovContra/Downloads/Opioid-Morphine-EQ-Conversion-Factors-Aug-2017.pdf.

C.M. Renier M. Juusola P. Hitz W. Steffen M.J. Asmus T. Craig J. Mardekian E.T. Masters T.E. Elliott Haller, I.V. Enhancing risk assessment in patients receiving chronic opioid analgesic therapy using natural language processing, October 2017. URL https://www.ncbi.nlm.nih.gov/pubmed/28034982.

Health and Human Services. The prescription drug and heroin overdose epidemic, March 2019. URL https://www.hhs.gov/opioids/.

H. Jalal C. Chang G. Cochran J. Donohue Lobo, C. Panel paper: Using unsupervised machine learning to identify potentially problematic opioid use in medicare, November 2017. URL https://appam.confex.com/appam/2017/webprogram/Paper23485.html.

NIDA National Institute On Drug Abuse. Overdose death rates, Jan 2019a. URL https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates.

NIDA National Institute On Drug Abuse. Opioid overdose crisis, Jan 2019b. URL https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis.

National Institute on Drug Abuse. Opioids, 2019. URL https://www.drugabuse.gov/drugs-abuse/opioids.

J. Borodovsky P. Mateu-Gelabert H. Guarino Sage Crosier, B. Finding a needle in the haystack: Using machine-learning to predict overdose in opioid users, February 2017. URL Findinganeedleinthehaystack:Usingmachine-learningtopredictoverdoseinopioidusers.

L.R. Webster and R.M Webster. Predicting aberrant behaviors in opiod-treated patients: Preliminary validation of the opiod risk tool, December 2005. URL https://www.ncbi.nlm.nih.gov/pubmed/16336480.