
As per the National Institute of Drug Abuse, more Americans die every day from opioid overdose. The problem of obtaining opioids through non-illicit prescriptions (obtained through doctor-patient relationship) is rarely discussed. This project analyses the increase in opioid recommendation in the U.S. population by analyzing the datasets provided by Medical Expenditure Panel Survey (MEPS). The survey data is analyzed for patterns and documented in the form of workflow. This project also aims at creating a predictive model with independent variables to identify the likelihood of opioid addiction.

Headings:

Predictive Modelling

Opioid Addiction

Healthcare Analytics

Workflow model
ANALYZING THE TRENDS IN PRESCRIPTION OPIOID FOR THE U.S. CIVILIAN NONINSTITUTIONALIZED POPULATION

by
Kirubanand Chandrasekaran

A Master’s paper submitted to the faculty of the School of Information and Library Science of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in Information Science.

Chapel Hill, North Carolina
April 2019

Approved by

Javed Mostafa
# TABLE OF CONTENTS

1. INTRODUCTION ......................................................................................................................... 2
  1.1 OPIOIDS ................................................................................................................................. 3
  1.2 ABUSE AND EPIDEMIC .......................................................................................................... 3

2. LITERATURE REVIEW ................................................................................................................ 6
  2.1 HISTORY OF OPIOID ABUSE AND IT’S GROWTH ............................................................... 6
  2.2 PEOPLE AFFECTED AND REMEDIAL EFFORTS BY ORGANIZATIONS ............................... 8
  2.3 DATA ANALYSIS AND OPIOID EPIDEMIC .......................................................................... 9
  2.4 PREDICTIVE ANALYTICS ON HEALTH RECORDS .............................................................. 11
  2.5 PREDICTIVE ANALYTICS AND OPIOID EPIDEMIC ............................................................ 13
  2.6 EHR WORKFLOW .................................................................................................................. 14

3. METHODS .................................................................................................................................... 16
  3.1 DATA ....................................................................................................................................... 18
  3.2 DATA CLEANING AND MANIPULATION .............................................................................. 19
  3.3 ANALYSIS ............................................................................................................................. 21
  3.4 PREDICTIVE MODELLING ...................................................................................................... 26

4. RESULTS AND DISCUSSION ..................................................................................................... 34

5. CONCLUSION AND FUTURE WORK ....................................................................................... 38

BIBLIOGRAPHY .............................................................................................................................. 40

APPENDIX .......................................................................................................................................... 44
1. Introduction

Opioid abuse is an increasing problem in the U.S. On February 2018, President Trump vowed to end the country’s “terrible drug epidemic.” The tone says a lot about the problem of increasing opioids in the U.S. In 2015 more than 52,000 Americans dies of overdoses. Of those approximately 33,000 of them where fatal overdoses from opioids, including prescription painkillers and heroin. The deaths from opioids are more in the big cities, but you could observe a pattern in the rural Appalachia, North East (NE) and the Midwest. The economic status of the victims is mainly middle-class people located in suburbs and rural towns.

Fig 1. Opioid overuse deaths across the U.S.
src: Centers for Disease Control and Prevention
The situation of opioid abuse is adverse, while the government is trying to fight the epidemic on a war footing with the help of federal agencies, this project aims at making an exploratory analysis and introduce insights in the opioid usage from the Medical Expenditure Panel Survey (MEPS) data, a rich source of medical expenditure survey across the user with anonymous users. Later, I applied predictive machine learning algorithms to identify factors influencing the trend in the opioid usage increase.

1.1 Opioids

Opioids are the class of drugs naturally found in the opium poppy plant. Some prescription opioids are made from the plan directly, and others are made by scientists in labs using the same chemical structure. Opioids are often used as medicines because they contain chemicals that relax the body and can relieve pain. Prescription opioids are used mostly to treat moderate to severe pain, though some opioids can be used to treat coughing and diarrhea. Opioids can also make people feel very relaxed and "high" - which is why they are sometimes used for non-medical reasons. This can be dangerous because opioids can be highly addictive, and overdoses and death are common. About 80 percent of people who use heroin first misused prescription opioids.

1.2 Abuse and Epidemic

Opioids range from medically prescribed pain relievers to the illegally obtained heroin. These drugs interact with the opioid receptors on nerve cells in the body and brain. Opioid pain relievers are safe when take over a short period of time as prescribed by the physician. Prescription opioids cause the risk of dependence and addiction. Abuse can be defined as taking too much of medicines or taking someone else’s prescribed medicine.
The abuse of opioids leads to addiction including prescription pain relievers, heroin, and synthetic opioids. An estimated 52 million people in the U.S. have used prescription medicine for non-medical reasons in their lifetimes. Everyday 130 people in the United States die after over dosing on opioids. Opioid misuse alone in the United States is causing $78.5 billion a year, including costs of healthcare, lost productivity, addiction treatment, and criminal justice involvement.\(^3\)

The opioid crisis has its roots back to the 1990’s when the pharmaceutical companies reassured the medical community that patients would not become addicted to the prescription opioid pain relievers, and healthcare providers began to prescribe them at greater rates. This subsequently led to widespread diversion and misuse of these medications before it became clear that these medications could be highly addictive. As a result, the opioid overdose results began to increase. In 2017, more than 47,000 Americans died as a result of an opioid overdoses, including prescription opioids, heroin, and illicitly manufactured fentanyl, a powerful synthetic opioid.\(^3\) Some facts about the opioid epidemic are, roughly 21 to 29 percent of the patients prescribed opioids for chronic pain misuse them.\(^4\) Opioid overdoses in large cities increase by 54 percent in 16 states. This issue has become public health crisis with devastating consequences including increases in opioid misuse and related overdoses, as well as the rising incidence of neonatal abstinence syndrome due to opioid use and misuse during pregnancy. The increase in injection drug use has also contributed to the spread of the infectious diseases including Human Immuno deficiency Virus (HIV) and hepatitis C.
The history repeats itself, science can be an important solution in resolving this crisis. Data analysis related to the opioid crisis reveals a pattern of opioid prescription including dose and duration with the patient’s risk factors of age, gender and condition are major indicators to determine whether a patient will become opioid addict.

With the aid from various previous research results and the unexplored MEPS data, this project aims at getting more insights on the opioid epidemic as well as lay a foundation for future studies. The literature review describes the technical aspects of the project including the methods, theories and tools used to arrive at the results.
2. Literature review

This chapter walks through the previous research on opioid crisis and the data-driven preventive measures taken to combat the epidemic. This chapter introduces brief discussion about the preventive measure taken by organizations and the use of data analysis in helping them. The opioid abuse history and importance is discussed to give a context to the user followed by the ways of identifying population affected by the opioid crisis.

2.1 History of Opioid Abuse and it’s Growth

The opioid addiction epidemic is increasingly becoming a big crisis in America. As per the Centers for Disease Control and Prevention (CDC), opioids (including prescription opioids and heroin) kill more than 33,000 people annually. The life expectancy in the U.S has dropped in an alarming rate, for men it fell down by a value of two-tenths of a year -- from 76.5 to 76.3. For women, it dropped one-tenth – from 81.3 to 81.2 years. Experts believe that this drop of life expectancy is unexpected. According to the Substance Abuse and Mental Health Services Administration (SAMHSA) Survey, nonmedical use of prescription drugs is the second most prevalent type of drug abuse, after marijuana.

From 1999-2017, almost 400,000 people have died from an overdoes involving any opioid, including prescription and illicit opioids. The opioid overdose death can be classified into three distinct waves as depicted in the below figure.
The first wave began with increased prescribing of opioids in the 1990s, with overdose deaths involving prescription opioids both natural and synthetic opioids and methadone till 1999. The second wave began in 2010, with rapid increases in overdose deaths involving heroin. The third wave began in 2013, with significant increases in overdoses deaths involving synthetic opioids – particularly those involving illicitly-manufactured fentanyl (IMF). The IMF market continues to change, and IMF can be found in combination with heroin, counterfeit pills and cocaine. Now that we have regulations over the illegally obtained opioids, the prescription opioid is on the rise.
2.2 People affected and remedial efforts by organizations

The opioid users’ characteristics differ widely in the U.S. from 2000 to 2010. The population is more likely older and receive public health insurance, more likely to be unemployed, more likely to have higher education, more likely to be obese, and in a low-income category. In 2010, a higher proportion of users were older and had a higher proportion who were non-Hispanic white, female, covered by public insurance, lower income category. In addition, they all reported poorer mental health.\textsuperscript{10}

According to the world drug report released by the UN Office on Drugs and Crime (UNDOC) 2016, the International Narcotics Control Board (INCB), based in Vienna, Austria, noted with great concern, the largescale opioid, prescription drug and heroin abuse problem that continue to affect the United States, claiming tens of thousands of victims each year.\textsuperscript{11} The report also acknowledges the approval of the Drug Enforcement Administration’s comprehensive action plan to address opioid addiction and the allocation by the Government of $27.6 billion for the 2016 fiscal year to aid the implementation of 2015 National Drug Control Strategy. The Food and Drug Administration (FDA), also joined forces by releasing an action plan against opioids in February 2016, in response to the ongoing crisis of opioid abuse, dependence and overdose in the United States. A class-wide safety labelling changes for immediate-release opioid pain medications have been announced.

Health Insurance giants such as Blue Shield and retail pharmacies such as Walmart and CVS are also active participants to eradicate the epidemic.

On 22 July 2016, the Comprehensive Addiction and Recovery Act \textsuperscript{12} was enacted by the 114\textsuperscript{th} Congress to address the opioid crisis by authoring the U.S. department of
Justice to award grants to state, local and tribal governments to provide opioid abuse services, directs the Department of Veteran Affairs to expand its opioid safety initiative and help communities to develop and treat the overdoses problem. And it also provides exemptions from criminal and civil liability for those administering an opioid overdoses reversal drug or who contact emergency services in response to an overdose.

August 2017, Jeff Sessions ordered the launch of an Opioid Fraud and Abuse Detection Unit\textsuperscript{13} within the Department of Justice. The unit’s mission is to arrest individuals who commit opioid-related fraud. There has been a country wide regulation across clinics to control the problem.

2.3 Data Analysis and Opioid Epidemic

The opioid epidemic is a problem that’s analogous with the technology, with the increasing data and opioid usage and the technology to analyze them we can come up with interesting insights to mitigate the problem. The main data sources include large hospital systems and public health agencies to better analyze the possible outcomes.

The SAS institute conducted a study\textsuperscript{14} to identify that better data and analytics can improve better treatment protocols, both for pain in the first place and for remediation when patients are becoming dependent on drugs. To fight the opioid epidemic the physicians and their patients, medical policy makers and licensing boards, pharmaceutical companies and pharmacies all must work together and achieve the fundamental objectives of reducing addiction and deaths. The report indicates that a lot of effort has been put to Prescription Drug Monitoring Programs (PDMP), databases of prescriptions written and filled plus online portals for accessing the data. In the aggregate, these systems have a difference in finding and shutting down the “pill mills”. The Electronic Health Record (EHR) and
emergency room records needs to be compared. When a patient is visiting more than one specialist and getting more than one prescription, tracking them can be very hard. Prescribing physicians have access to data that is useful but incomplete.

Information is fragmented but still available, The CDC has issued guidelines about revealing chronic pain patients after three months, and Schedule 2 drugs. But many patients subtly demand their physicians for opioids and get prescribed by the system in unsafe quantities for long periods of time. Analytics can help the physician recognize the patient scenarios, prescribe correctly, and focus on best overall outcome. Larger organizations can do better in terms of handling opioid crisis with data analysis as they have all the related information like patient visits, physician remark, ER visits, prescriptions, test results in one place. Analytics can inform treatment guidelines, educational initiatives, and resource allocations, including treatment centers and community prescription drug take-back programs. These organizations can perform better to influence policy makers with actual proof.

The pharmacies need a methodology to compare the prescription cases that they receive and process their location, payment source, provider and patient mix. Data analysis can help them in identifying the anomalies and alert the authorities. Getting the data together helps the various constituencies work together. It helps them see and understand a better picture, as well as concentrate on the outcome they want. Reducing the spread of illegal street drugs would take a massive enforcement strategy.

A workflow consists of an orchestrated and repeatable pattern of business activity enabled by the systematic organization of resources into processes that transform materials, provide services, or process information. Using machine learning algorithms on the raw
data form the data science workflow. Any approach using data science should start from the root of the problem, which is the prescription for opioids. We need to identify the patterns and identify the anomalies in it to devise data science workflows for better outcomes in the future.

The CDC has awarded 28.6 million USD for big data analytics to track the abuse of opioids.\textsuperscript{15} With the new funding apportioned to 44 states and the District of Columbia, the CDC will use three new programs to enhance public health officials’ abilities to track, monitor, prevent and treat individuals experiencing problems related to substance abuse. Some of the activities include the use of prescription drug monitoring programs (PDMPs), messaging to communities about the impact of opioid abuse, and additional reviews of patient datasets, such as overdoes as fatality data, to get better target preventive care efforts and patient management programs.\textsuperscript{15}

2.4 Predictive Analytics on Health Records

Instead of simply representing information about past events to user, predictive analytics estimate the likelihood of a future outcome based on patterns in the historical data. This allows physicians, experts and administrative staff about potential events before they happen, and therefore make more informed choices about how to proceed with a decision.

This may seem fruitful but only few high-value cases exist throughout the medical ecosystem and may not always involve getting real-time alerts that require a team to immediately spring into action.\textsuperscript{17} Some of the areas where predictive analytics can be useful in the health ecosystem are: risk scoring, forecasting before patient, and managing the supply chain.
Risk scoring

Prediction and prevention always go together, perhaps more closely in the medical field. Organizations that can identify individuals with elevated risks of developing chronic conditions as early in the disease’s development will have a chance of treating themselves right. Risk scores can be created based on lab tests from patient’s health data. This modelling can be helpful to proactively identify patients who are at the highest risk of poor health outcomes and will benefit the most from the intervention is one solution believed to improve risk management for providers transitioning to value-based payment.

Forecasting before patient deterioration

While still in the hospital, patients face a number of risks for their well-being, the recovery rate may be very slow, or they may get infected by more hard-to-treat infection, or a sudden downturn due to existing clinical conditions. UPenn uses a machine learning EHRs to target server sepsis, the machine learning algorithm that continuously monitors EHR data for signs of sever sepsis can reduce the time to detection by 12 hours.

Managing the supply chain

The supply chain is one of the major puzzles to solve for every healthcare system, it is one of the tricky areas where the organizations can save their spending and improve efficiency. Predictive analytics are in demand among hospitals looking to reduce variation and gain more actionable insights into ordering patterns and supply utilization. Though it is very useful, only 17% of the hospitals use an automated healthcare supply chain management solution for inventory management. The Global Healthcare Exchange ranked predictive analytics for supply chain management as the number one item on the
executive wish list—a follow-up survey in 2018 found that adopting predictive analytics could remain a top priority.

Using analytics for supply chain management could save hospitals almost $10 million per year.\textsuperscript{19} Analytics also help in negotiating prices and reduce the unnecessary supplies and optimizing the ordering process.

Apart from the above mentioned uses of predictive analytics, some of the more notable areas are developing precision medicine and new therapies for patients and bolstering patient engagement and satisfaction.

\section*{2.5 Predictive Analytics and Opioid Epidemic}

The Mckinsey and Company has successfully conducted a data analysis on the opioid epidemic and came up with 10 insights through claim analysis.\textsuperscript{20} These insights are published to trigger further research and discussions. Some of the striking revelation from analysis are:

\begin{itemize}
\item Prescribing patterns are influenced significantly geography, even among patients undergoing similar type of care.
\item Providers frequently prescribe opioids to patients with known or potential risk factors for abuse.
\item Patients with concurrent prescriptions for an opioid and a behavioral health condition appear to have a 30\% or greater likelihood of developing future opioid dependence.
\item A small portion of opioid use originates in emergency departments.
\end{itemize}
Building a predictive model requires careful speculation of what specifically needs predicting, as this impacts the type of models and methods used later on. One common goal in predictive modeling is to accurately predict an outcome value for a new set of observations. This goal is known in predictive analytics as prediction (for a numerical outcome) or classification (for a categorical outcome). A different goal, when the outcome is categorical (e.g., adopter/non-adopter), is to rank a new set of observations according to their probability of belonging to a certain class.27

2.6 EHR Workflow

Electronic health records will eventually eliminate paper charts while bolstering the information today’s charts contain, making them more accessible, accurate and complete. Workflow is the lifeblood of a practice. And the efficiency of that workflow can affect a practice's quality of patient care and level of profitability. An EHR workflow that is built around proven processes and procedures can rapidly improve the proficiency and effectiveness of a practice. Workflows in EHR systems help us to analyze the procedures followed in a hospital and other medical facilities. Paper charts don’t empower patients the way EHRs do. Implementing workflow in EHR systems will make the patients to contribute directly to the data collection process. Some of the advantages of using an EHR workflow system are 1) Security and privacy of health information. 2) User interfaces that support clinical reasoning and decision making. 3) Shared application and network architectures. 4) Secondary use of EHR data to improve health.
The EHR workflow system involves a lot of data in short, big data. When talking in terms of big data, the biomedical scientists are facing new challenges of storing managing and analyzing massive amounts of datasets. There is a need for architecture to manage large amounts of EHR. Though the scope of this master’s project is a single use case. It’s aimed at documenting every single step in the process to create a better understanding.
3. Methods

The project has two main goals, extract data from MEPS and use it for getting insights on opioid abuse and document the process of obtaining the insights in form of workflow modules to help students understand the process in the future. The following section explains about the source of the data set, algorithms used for analysis and the technologies used to devise the workflow.
Fig 4. Workflow of the data science process

**Jupyter notebook**

In my case, I am hosting the application in the cloud itself with the data. There are clear boundaries between the application code and information storage. It gives users access to computational environments and resources without burdening the users with installation and maintenance tasks. Users - including students, researchers, and data scientists - can get their work done in their own workspaces on shared resources which can be managed efficiently by system administrators.

**Python**

Python is the programming language framework I use in the project to get things done. Why choose python, there are a number of reasons to it. The primary reason is being it is easy to understand for people. Since the primary objective of the project is to help people learn the data analysis process using jupyter hubs. Python code is easily readable for the students and hence they can make their own changes if they need for learning.

Python has many open source machine learning libraries that can be used to develop insights for data analysis.

**Pandas**

Pandas is a Python package providing fast, flexible, and expressive data structures designed to make working with structured (tabular, multidimensional, potentially heterogeneous) and time series data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real world data analysis in Python.
Additionally, it has the broader goal of becoming the most powerful and flexible open source data analysis / manipulation tool available in any language. It is already well on its way toward this goal.

3.1 Data

MEPS

The Medical Expenditure Panel Survey (MEPS), which began in the year 1996, is large-scale surveys of families and individuals, their medical providers (doctors, hospitals, pharmacies, etc.), and employers across the United States. MEPS collects data on the specific health services that America uses, how frequently they are used, the cost of these services, and how they are paid for, as well as data, scope and breadth of health insurance held by and available to U.S workers.21

For the analysis I use different data files from the MEPS data repository. To understand the kind of data needed for analysis lets re-iterate the problem statement, to analyze the trends in prescription opioids that are used among the U.S population. With just the prescription only the type of drug and their dosage is obtained. To add more details, information about the patient to get an overview about their ethnicity, employment, income group is needed. With this context, I collected the following data files from MEPS for a given year.

**Prescription records:**

This public use file provides detailed information on household-reported prescribed medicines for a nationally representative sample of the civilian noninstitutionalized population of the United States and can be used to make estimates of prescribed medicine utilization and expenditures for calendar year.22
Opioids list file

This file was sourced from a Kaggle challenge on U.S Opiate prescription overdoses, contains the names of all opioid drugs included in the data and overdoses.csv that contains information on opioid related drug overdose fatalities.23

Consolidated Data file

This file consolidates the person level details in a single file, some of the notable features that can be obtained file include language of interview variable, demographics, patient identifiers, health status, disability days variables, access to care, employment, quality of care, patient satisfaction, health insurance, and use variables.

3.2 Data Cleaning and Manipulation

Data obtained usually is raw and may contain duplicates. Data cleaning is the process of detecting and correcting the inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying or deleting the coarse data. Data cleaning can be performed by data wrangler tools like Open refine, or through scripting using python.

The first step in data cleaning is to read the files in suitable format. The MEPS data files are by default in the. spss format, which is the SAS transport format, to convert them to .csv which is suitable format for data analysis using python, I wrote a script to convert the files to readable .csv format.

```
~/Documents/OneDrive - University of North Carolina at Chapel H
r_project/data/2012 > python -m xport h138.ssp > h138.csv
```

Fig 5: Exporting SAS transport format data into csv
Prescribed medicines file

Step 1: Load prescribed medicines csv file into a pandas data frame

The csv frames are loaded into the online jupyterhub\(^4\) for faster processing and so that they can be embedded in the workflow system of analysis.

Step 2: Filter only required features

Since the data frame consists of more than 60+ features but for our analysis I considered only important features like drug name, drug id, prescribed medicine name, cost, consumption form, patient id.

Step 3: Filter only opioids

The file contains all prescribed medicines for the survey year, the focus is only interested in the opioid prescriptions and their associated data for the analysis. So, I applied another filter to match rows that contain opioids matching our common opioid list. The column ‘RXDRGNAM’ contains the common name for the drug, filtering this column gives the list of opioid only prescriptions from the dataset.

Step 4: Finding repeated opioid prescriptions

The data frame obtained from step 3 contains multiple prescriptions for a single patient identified by their patient ID (DUPERSID). The patient ID is considered as the primary key for the data frame so, the data frame is aggregated to add two more columns for each patient:

- ‘RXXPTOTAL’ - indicates the total amount spent by each patient on opioid prescriptions.
- ‘RXTIMES’ – indicates the number of times the patient has taken opioid prescriptions.
The consolidated data file as described above contains all the details about a patient like race, age, income group, health status. Since the prescribed medicine data frame is consolidated to represent each patient by a row, this makes it easier for us to compare with the consolidated data file to get the details of the patient. The consolidated data file is merged with the cleaned prescribed medicine opioid file based on the primary key DUPERSID, which identifies each user in both the data frames. The overall super data frame that contains all the required information for analysis for all opioid prescribed patients of that year. The same process is repeated for all years from 2010 to 2014 to get the 5 opioid prescribed patient data frames.

3.3 Analysis

To identify why the problem was chosen in the first place, some analysis on the data before identifying predictive modelling usages is needed. I filtered opioid only patients as it contains both patients with single use and repeated offenders. As the focus is only on repeated offenders some exploratory analysis on the data to identify these patients is needed.
Exploratory Analysis:

Exploratory Data Analysis refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations.\textsuperscript{25}

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure7}
\caption{Percentage of opioid prescription in MEPS data}
\end{figure}

The above figure implies a sharp decrease in the opioid prescriptions for the year 2014 and 2015 indicates efforts towards decreased opioid prescriptions by government organizations. To explore further I worked on the expenditure towards opioid prescriptions for each year to find if the spending has increase. I calculated the average amount spend on opioids by patients for the years which turns out to be the highest on the year 2015 with an average spending of $308. Another reason to justify the need for the study.
The data obtained so far is divided into data frames for each year, to create a model which can predict if the user is opioid addicted or not. Further cleaning of the data and combination of a larger data set with categorical variables to devise the model is required. The steps involved are explained below:

**Step 1: Removing irrelevant features:**

The features which are considered to be less important by the definition of the problem are ignored in this step for each of the data frame from the years 2010 to 2015. As I am trying to predict if the user can be a potential addict or not, So I am ignoring the following features:

- Total expense and expense related (TOTEXP10) – the cost of each opioids varies, some are very expensive, and some are commonly affordable.

- Opioid type (RXNDC, RXQUANTITY, RXFORM, RXSTRENG) – The data frames consist only of only opioid users, the task is to differentiate between opioid abusers and people who consume for medical conditions. So, the names cannot
particularly differentiate, when all opioids can be dangerous when consumed above prescribed limit.

**Step 2: Rename columns to make the data frame consistent for appending:**

Since the MEPS survey methods change from year to year, some of the columns have different feature names over the year with the same description. These columns are identified and renamed.

- RACEX to RACETHX
- HIGHTED to EDUYRDG

**Step 3: Appending the data frames**

The data frames from each year are appended to form a larger super data set containing more variables.

```python
bigdata = pd.concat([df_p_2010, df_p_2011, df_p_2012, df_p_2013, df_p_2014, df_p_2015], ignore_index=True)
```

**Fig 9. Concatenating yearly data frames**

**Step 4: Cleaning individual features in Openrefine**

Some of the other changes made to the variables include calculating the age of the patient at the time of the survey, cleaning invalid values under the categorical variables, forming a super key (SUPERSID) by combining the patient ID (DUPERSID) and Year (YEAR) as many of the patients are attend the survey twice in different years.

**Fig 10. Repeated patient**
The above figure illustrates that a particular patient has taken the MEPS survey in consecutive years, but his characteristics have changed over the time.

**Step 5: Determining addiction in patients:**

A simple criterion to differentiate normal opioid users from the repeated users is to find the number of prescriptions each patient has obtained. I determined a predicted variable ‘ADDICTED’ which is binary. If the patient has more than 4 visits which implies 4 prescriptions, then the patient is marked as opioid addict. This claim is a crucial part of the thesis from the basis of the CDC data for prescribing pattern of opioids. More than 17% of Americans had at least one opioid prescription filled, with an average of 3.4 opioid prescriptions dispensed per patient.26

The number of visits is not included in the process of building the final model as it affects the performance of the model entirely.

A brief explanation of the features in the above figure:

- **RACETHX** – Race of the patient
- **RTHLTH53** – Perceived health status reported by the patient
- **TTLPX** – Total expenses on opioids on that year by the patient
- **INSCOV** – Insurance coverage type of the patient
• ADSMOK42 – Smoking habit at the time of the survey
• BMINDX53 – BMI index of the patient
• ADHOPE42 – How often does the patient feel hopeless
• ADIWN42 – How often does the patient feel rejected or down
• ADDPRS42 – How often does the patient feel depressed
• HIDEG – Highest degree at the time of the survey
• JTPAIN53 – Reported if there are any joint pains at the time of the survey.

The reason behind the selection of these features is to consider the social, medical and educational factors into consideration when classifying a patient.

### 3.4 Predictive Modelling

I created the data set with a target binary variable with supporting categorical variables to form a predictive model. The important aspect of predictive modelling is the time spent in finding the right model suitable for the problem we are trying to solve. For our problem of building a model to predict the outcome variable or target variable we need to pick out categorical variables and see their impact on the target variable and verify. For the process of predictive modelling we use scikit-learn, a machine learning package for python. Scikit-learn is a Python module integrating a wide range of state-of-the-art machine learning algorithms for medium-scale supervised and unsupervised problems. This package focuses on bringing machine learning to non-specialists using a general-purpose high-level language. Emphasis is put on ease of use, performance, documentation, and API consistency. It has minimal dependencies and is distributed under the simplified BSD license, encouraging its use in both academic and commercial settings.\(^\text{28}\)
The impact of variables is analyzed by plotting them in a graph against the predictor variable. A value of ‘1’ indicates the person is opioid addicted and ‘0’ indicates the person is not addicted implying they received less than 4 prescriptions that year. The results can be seen below:

Fig 12. Smoking vs No. of opioid addicted patients

Fig 13. Race vs No. of opioid addicted patients
The above variables indicate that they affect the outcome of the target variable. The number of patients changes significantly for each value of the variable considered. These may also happen by random chances or the data may be biased. So, Further exploration is required in the process of building the models. Above that I also included numerical variables like age, BMI and total expenses.

**Decision Trees:**

A decision tree is a flowchart-like tree structure where an internal node represents feature (or attribute), the branch represents a decision rule, and each leaf node represents
the outcome. The topmost node in a decision tree is known as the root node. It learns to partition on the basis of the attribute value. It partitions the tree in recursively manner call recursive partitioning. This flowchart-like structure helps you in decision making. It's visualization like a flowchart diagram which easily mimics the human level thinking. That is why decision trees are easy to understand and interpret.29

The next step in decision tree is to select the features for the independent and target variables. The next process is to divide the data to training and test data set, I chose 70% training for all the analyses and 30% test data for verifying the model. The features are passed on to the decision tree and the accuracy is verified.

The initial accuracy turns out to be 66% percent which is a marginally good model but to explore the reasons for the low accuracy and to improve the performance of the model, I analyzed the target variable. It is found that the number of opioid addicted patients in the dataset is very low and the pruning is set to null resulting in an over-fitting model. To address these problems, I resorted to logistic regression.

![Fig 16. Target variable imbalance](image-url)
Logistic Regression

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. I performed a binary logistic regression on the data.\textsuperscript{30}

From the observations so far, I identified that the data is imbalanced, and there are number of independent variables to predict the target variable. To solve the problem of imbalance data of the target variable a two-step process to enrich the data and actual model building is performed.

Binary Encoding

The independent variables which are categorical are picked and a new column for each category is created using the python function – getdummies().

```
Out[38]: array(['SUPERSID', 'RACETHX', 'SEX', 'RTHLTH53', 'TLTPX', 'INSCOV',
                  'ADSNOK2', 'BMINDX53', 'ADHOPE42', 'ADDOWN42', 'ADDPRS42',
                  'ADPAIN42', 'KIDE', 'JTPAIN53', 'MILDIF53', 'AGE', 'RKTIMES',
                  'ADDICTED'], dtype=object)
```

*Fig 17. Columns before binary encoding*
The problem of under sampling of target variable still exists, to overcome this problem I use the SMOTE (Synthetic Minority Oversampling Technique) algorithm.

**SMOTE:**

An approach to the construction of classifiers from imbalanced datasets is described. A dataset is imbalanced if the classification categories are not approximately equally represented. Often real-world data sets are predominately composed of “normal” examples with only a small percentage of “abnormal” or “interesting” examples. It is also the case that the cost of misclassifying an abnormal (interesting) example as a normal example is often much higher than the cost of the reverse error. Works by creating synthetic samples from the minor class (addiction) instead of creating copies. Randomly choosing one of the k-nearest-neighbors and using it to create a similar, but randomly tweaked, new observations.
Fig 18. Oversampled data with balanced target variable

Recursive Feature Elimination

After binary encoding I ended up with 82 variables in total including the numerical variables which were not converted. The goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the importance of each feature is obtained either through a coef_attribute or through a feature_importances_attribute. Then, the least important features are pruned from the current set of features. That procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached.33

Fig 19. Result of RFE

The result of running the feature ranking algorithm is that I got rid of features not leading us to implement the actual model. On running the actual model, the p values of

```
most of the variables above 0.05 so they are eliminated including the race category which does not seem to have enough data. In terms of logistic regression, they are pruned.
4. Results and Discussion

The results can be divided into two sections for the two algorithms.

Decision Tree

The decision tree initially had an accuracy of 66% with the unbalanced data, with the process of applying binary encoding and using the SMOTES algorithm to balance the sampling of the target variable, I applied pruning to the decision tree by restricting the level by 4. The resulting accuracy is 74%. The resulting decision tree looks like this:

Some of the insights that can be obtained from the decision tree are, most of the opioid addicts lie above 38.5 years of age. The patients with repeated opioid use had a perceived health status between fair and poor. This indicates that most lied to the physician to get an opioid prescription repeatedly. Though this model may seemingly
accurate, it can be noticed that it does not take other features into consideration, leaving us with a doubt of under-fitting.

**Logistic Regression**

From the last step, the features with a p value greater than 0.05 are removed leaving us with only the age and perceived health. This result supports the output of the decision tree that other features like race, smoking habits does not have considerable effects on the outcome of the target variable.

```python
y_pred = logreg.predict(X_test)
print('Accuracy of logistic regression classifier on test set: {:.2f}'.format(logreg.score(X_test, y_test)))

from sklearn.metrics import confusion_matrix
confusion_matrix = confusion_matrix(y_test, y_pred)
print(confusion_matrix)
```

<table>
<thead>
<tr>
<th></th>
<th>Predicted No</th>
<th>Predicted Yes</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual No</td>
<td>747</td>
<td>373</td>
<td>1120</td>
</tr>
<tr>
<td>Actual Yes</td>
<td>229</td>
<td>858</td>
<td>1087</td>
</tr>
<tr>
<td>Total</td>
<td>976</td>
<td>1231</td>
<td>2207</td>
</tr>
</tbody>
</table>

*Fig 20. Accuracy and confusion matrix of the model*

The accuracy has slightly improved over the original decision tree model and the confusion matrix indicates 1120 predictions right and 1087 predictions wrong from the training and test data split.

AUC - ROC curve is a performance measurement for classification problem at various thresholds settings. ROC is a probability curve and AUC represent degree or
measure of separability. It tells how much model is capable of distinguishing between classes. Higher the AUC, better the model is at predicting 0s as 0s and 1s as 1s. By analogy, Higher the AUC, better the model is at distinguishing between patients with addiction and no addiction.\textsuperscript{34}

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{Receiver_operating_characteristic.png}
\caption{ROC curve}
\end{figure}

A good classifier always stays away from the purely random classifier line. The farther away and closer to the top-left corner, the better the model. The accuracy of the model has improved to 73%.

<table>
<thead>
<tr>
<th>precision</th>
<th>recall</th>
<th>f1-score</th>
<th>support</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.77</td>
<td>0.67</td>
<td>0.71</td>
</tr>
<tr>
<td>1</td>
<td>0.70</td>
<td>0.79</td>
<td>0.74</td>
</tr>
</tbody>
</table>

\begin{figure}[h]
\centering
\includegraphics[width=0.5\textwidth]{Precision_and_Recall.png}
\caption{Precision and Recall}
\end{figure}

Precision is ability of the classifier to not label a sample as a positive even though it is negative. Recall is ability to find all positive samples of the classifier. From the results its inferred that the model is better in finding the addicted patients (0.79) than finding a non-addicted patient (0.67). Support indicates the number of occurrences of each classifier in the test data set.
Other interesting insight is the identification of most opioids used by the patients:

- ACETAMINOPHEN-CODEINE
- BUPRENORPHINE-NALOXONE
- CARISOPRODOL
- CODEINE
- FENTANYL
- HYDROCODONE
- HYDROMORPHONE
- MORPHINE
- OXYCODONE
- TRAMADOL

It has been found that there is a wide spread increase of prescription opioids over the years and the spending has also considerably increased every year. However, there is a drop in the expenses in the year 2014 indicating a sampling variation. The finding about the most common opioids agrees with the publicly available information by the American addiction center.35
5. Conclusion and Future Work

Contribution

The thesis aimed at following the entire data science process from collecting raw data to generating models to infer information. The entire process workflow is documented to help further research on this front.

The logistic regression resulted in selecting the influential set of variables in determining if a patient is opioid addict or not. These features include the final set of variables that helped in building the regression model are – Age, race, public health insurance coverage, depressed for more than half of the days, pain affected their normal functioning and joint pain. The factors like education, total expenses did not have an impact on the final model.

Implication

The logistic regression model devised above has identified some independent variables that affect the prescription opioid analgesic use. This information can be useful when we are targeting initiatives to control the spread of opioid use and to identify the people group. The exploratory analysis revealed that most of the overuse belonged to non-Hispanic white, female population. Even though this study produces some interesting insights, the results should be handled with caution for future research due to the limitations of MEPS data.
Limitation

MEPS data is restricted to non-institutionalized Americans, this does not include military population where the problems of opioid abuse are widespread after retirement from active service. In fact, a part of the MEPS survey is self-reporting of the information by the patient. The patient cannot be trusted in the self-reporting model especially when we are dealing with overdoses problems.

Future work

Our model does not include the factors of cost, form and dosage of consumption of the opioids. If this information along with the deeper understanding of the health variables present the consolidated data file for each patient can give more data points for future research on opioid addiction. Increasing the accuracy of the models is always a problem of data science. With this model being an above average accurate one, future models with better prediction accuracy can be path-breaking.
Bibliography


[19] LaPoint, J. (February 16, 2017). 78% of Hospital Staff Still Face Manual Supply Chain Management. Retrieved from: 


https://meps.ahrq.gov/mepsweb/

https://meps.ahrq.gov/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-188A


https://jupyter.org/

https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

[26] Changes in Opioid Prescribing Practices. Retrieved from: 
https://www.cdc.gov/drugoverdose/data/prescribing.html


[29] Navlani, Avinash. (December, 2018). Decision Tree Classification in Python. Retrieved from:  
https://www.datacamp.com/community/tutorials/decision-tree-classification-python

https://www.statisticssolutions.com/what-is-logistic-regression/


[33] sklearn.feature_selection.RFE. Retrieved from:  

[34] Narkhede, Sarang. (June, 2018). Understanding AUC - ROC Curve. Retrieved from:  
https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

https://americanaddictioncenters.org/the-big-list-of-narcotic-drugs
Appendix

Workflow of cleaning raw data from MEPS

Execute the actions in top-down order (green on completion)

- Read CSV and filter required features
- Extract Opioid Prescriptions
- Calculate RX times
- Join with consolidated data file
- Write Data frame to CSV

Aggregating prescriptions for every patient

<table>
<thead>
<tr>
<th>DUPSID</th>
<th>DRUCIDX</th>
<th>LINKIDX</th>
<th>RXNAME</th>
<th>RXNDC</th>
<th>RXQUANTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10007102</td>
<td>10007102</td>
<td>10007102</td>
<td>10007102</td>
<td>CARISOPRODOLE 603258228</td>
</tr>
<tr>
<td>1</td>
<td>10030103</td>
<td>10030103</td>
<td>10030103</td>
<td>10030103</td>
<td>APAP/COCAINE 603233832</td>
</tr>
<tr>
<td>2</td>
<td>10035101</td>
<td>10035101</td>
<td>10035101</td>
<td>10035101</td>
<td>TRAMADOL HCL 378415101</td>
</tr>
<tr>
<td>3</td>
<td>10041101</td>
<td>10041101</td>
<td>10041101</td>
<td>10041101</td>
<td>TRAMADOL HCL 65152062710</td>
</tr>
<tr>
<td>4</td>
<td>10079102</td>
<td>10079102</td>
<td>10079102</td>
<td>10079102</td>
<td>APAP/COCAINE 93015010</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>RXFORM</th>
<th>RXFRMUNT</th>
<th>RXPT10X</th>
<th>RXSTRENG</th>
<th>RXPTOTAL</th>
<th>RXTIMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>TAB</td>
<td>10.00</td>
<td>350</td>
<td>10.00</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>TAB</td>
<td>3.00</td>
<td>360/30</td>
<td>3.00</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>TAB</td>
<td>6.00</td>
<td>50</td>
<td>6.00</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>TAB</td>
<td>2.55</td>
<td>50</td>
<td>2.55</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>TAB</td>
<td>5.45</td>
<td>9</td>
<td>5.45</td>
<td>1</td>
</tr>
</tbody>
</table>

Fig A1: First step of the data science workflow – Jupyterhub
Workflow of combining data from individual year

Execute the actions in top-down order (green on completion)

---

Head of individual dataframe after renaming features to align the dataframes

<table>
<thead>
<tr>
<th>DUIDSID</th>
<th>RACETHM</th>
<th>DOBYY</th>
<th>SEX</th>
<th>HTNLTH53</th>
<th>TTLPX</th>
<th>INSCOV</th>
<th>ADGNRK42</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.0</td>
<td>1.0</td>
<td>1934.0</td>
<td>2.0</td>
<td>4.0</td>
<td>14000.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>-1.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>-1.0</td>
</tr>
<tr>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>BMINDK53</th>
<th>ADDOPB42</th>
<th>ADDOWN42</th>
<th>ADDPRS42</th>
<th>ADPAIN42</th>
<th>MIDEA</th>
<th>JTPAIN53</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>25.7</td>
<td>2.0</td>
<td>2.0</td>
<td>3.0</td>
<td>4.0</td>
<td>3.0</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>8.0</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>8.0</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>8.0</td>
</tr>
<tr>
<td>1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>-1.0</td>
<td>8.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MILDIF53</th>
<th>YEAR</th>
<th>RTTIMES</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>3.0</td>
<td>2011</td>
</tr>
<tr>
<td>1.0</td>
<td>2011</td>
<td>1</td>
</tr>
<tr>
<td>1.0</td>
<td>2011</td>
<td>1</td>
</tr>
<tr>
<td>1.0</td>
<td>2011</td>
<td>7</td>
</tr>
<tr>
<td>1.0</td>
<td>2011</td>
<td>10</td>
</tr>
</tbody>
</table>

Fig A2: Second step of the data science workflow – Jupyterhub
Number of Opioid Prescribed Patients among Races

Race: Asian

YEAR: 2015

Education level distribution of opioid overuse patients

1-No Degree | 2-GED | 3-Highschool | 4-Bachelor Degree | 5-Master Degree
6-Doctorate | 7-Other | 8-Under 16 | 9,-1,-9 - Inapplicable

Fig A3: Inferences from the data science workflow – Jupyterhub
Number of Opioid overuse Patients among Races

1-Hispanic | 2-White | 3-Black | 4-Asian

![Chart](chart.jpg)

Number of Opioid Prescribed Patients among Races

Year: 2014  
Race: Asian  
AGE(tango): 24 – 68

Age distribution of opioid overuse patients

![Chart](chart2.jpg)

Fig A4: More Inferences from the data science workflow – Jupyterhub