GOPS: A Machine Learning Framework for Opioid Abuse Analysis

Shuo Liu
UNI: sl4921
sl4921@columbia.edu

Yu Li
UNI: yl4736
yl4736@columbia.edu

Yujin Chen
UNI: yc3851
yc3851@columbia.edu

Abstract

With the development of the medical industry, the opioids as common painkillers are being widely used in prescription therapy and recreation. The issue of opioid abuse has drawn public attention in recent years. To analyze and predict the abuse of drugs would be beneficial for people to better understand the upcoming crisis. Based on the existing previous work on similar topics, we explore the influence of social group features and the movement of the population as well the potential connection with heart diseases like coronary and stroke. We propose a framework that integrates various machine learning models to forecast and analyze this public safety problem, including to identify areas of high risk for future outbreaks and to find out the relation to several common heart diseases. Through our work, we attempt to help governments and medical institution to identify possible strategies to address the opioid crisis.

1. Introduction

The United States is in the grip of a nationwide opioid crisis, whether for the prescription purposes or for recreational purposes. The Centers for Disease Control and Prevention is working hard to sort out this crisis and avoid the harmful health repercussions from the epidemic, such as opioid use disorder, HIV infections, as well as some heart diseases [1]. The Federal Bureau of Investigation and Drug Enforcement Administration, among other Federal Agencies, are facing serious challenges to enforce current law systems to cater this crisis as well.

The opioid crisis would also lead to severe ramifications on different sections of the US economy. It would become difficult for businesses that need high-standard labor skills, sensitive work or security precautions, to have these positions filled if the crisis spreads to different sections of the US population. Furthermore, if the number of people addicted to opioids increases especially for those elderly people, it will become a headache for health cares and assisted living facilities. A mechanism is needed to trigger an alarm to the society and local governments about the incoming crisis. This would have significant economic and social values.

Previous research has indicated that drug usage is linked to geographic location and regional demographic features [2], and that it can induce heart diseases like coronary heart disease, heart attack, cardiac dysrhythmia, etc [3, 4]. If this tendency continues, it will have disastrous consequences for the country's economy and the health of its citizens. In this paper, we looked at the opioid problem in United States counties and try to figure out what are the reasons that causes it and how it's affecting people's health. Our targets are to develop mathematical models to predict the crisis' evolution and determine its causes and consequences. Specifically, we are going to solve the following problems.

How many opioid abuse cases are expected to be reported in the near future? Which social groups in terms of education experience, ethnicity, religious affiliation, and income, are more likely to use opioids, and which factors contribute to this growth? Intuitively, people's migration contributes a lot to the cultural integration and resident lifestyle, thus we also aim to figure out what role do population movement play in the opioid abuse crisis. Besides, what is the connection between heart diseases and opioid abuse?

Machine learning especially deep learning with neural work is often used to deal with prediction problems. Chen pointed out in an article about how he and his team applied machine learning in prediction of the expectations about medicine based on big data analysis [5]. Pham et al., on the other hand, proposed a deep learning approach and designed an end-to-end long short-term memory model named DeepCare to predict healthcare trajectories from medical records [6]. Both of the teams take advantage of this data analyzing techniques and deliver decent results and work, which inspires up to facilitate these useful techniques to predict the opioid crisis.

To solve this problems, we propose a Geographic Opioids Prediction base on Socio-demography (GOPS) framework, which predicts possible future hot spots in terms of various opioid types through graph structured analysis and

the corresponding possible diseases based on the their correlation. Our contributions are listed as below.

1. We use advanced machine learning models to make baseline predictions of the natural growth of different types of opioids.

2. We find the groups who are susceptible to opioids in terms of socio-demography. By extracting strong relevance features, the number of opioids based on the composition of the society is predicted. Based on this, we enhance this model by taking the migration rate into consideration.

3. We analyze the effects of different opioids on common heart diseases.

The rest of this paper is organized as follows. In the Sec. 2, we listed some related previous work and discussed their limitations. We introduce our GOPS framework in the Sec. 3, explaining the details of the construction of machine learning and deep learning models and discussing the rationality of these models. The experiment results of our models are shown in the Sec. 4, which illustrates the dataset we use, the platform and tools setup and the evaluation metrics. In the end, we discuss the vulnerability of our framework and the potential work in the future in the Sec. 5.

2. Related work

- **Opioids and Drug Abuse Analysis** Analysis of opioids and other drugs has long been a hot topic among data analysts. Back in 2009, Sebastiano et. al. used traditional statistical methods to estimate the risk of opioid misuse among adolescents [7]. Kendler et. al. found genetic and growth environment may play an importance role in drug abuse, which is one of the first well-developed work linking the problem of drug abuse to social problems [8]. With the development of machine learning technology, some researcher tried to exploit the huge potential of this modern statistic and prediction tool for analyzing the drug abuse. Ding et. al. made a review on similarity-based machine learning approaches for predicting drug abuse [9]. Although this work has similar goals to ours, it was first proposed at a time when deep learning predictive solutions were not yet available. In addition, with the development of medical service and entertainment industry, people are found to use new drugs, and it is no longer accurate to predict the future by using the past methods and data.

- **2019 The Mathematical Contest in Modeling (MCM) Problem C** Based on these studies of opioid growth model analysis, MCM 2019 has officially kicked off a heated discussion on this topic. There have been many groups that worked on this problem, and each analyses the problem with a different insight. In ShanghaiTech University group's work [10], the possible future drug usage was determined by a recommendation system. Support vector regression, which is suitable for multiple dimensional regression, is used to distinguish higher-risk counties. Principal component analysis feature extraction and association rule learning, which are used to discover the correlation between variables in a large database, were used for social features analysis. In the first group from the University of Colorado Boulder[11], an interesting gravity map is used to describe the growth of opioid cases between the states. In other works from the University of Colorado Boulder, Markovian assumption was used to reduce computational power needed [12], and random work is used to finely tune the model [13]. Though, our first two tasks are based on this competition and thus similar to these works, we expand these works of five states to all United States' counties. And we further analyzed the social impact of abuse of these drugs, which is not covered in these works. In addition, some works use PCA to conduct dimension reduction, which doesn't have interpretability. We use correlation analysis tricks to manually select some important ones to keep the interpretability of feature vectors. It's worthy to notice that, in all the models mentioned above, the geographic distance among counties and their social similarity played important roles in model training. This is reasonable as counties are more vulnerable if they are close to the center of opioid crisis outbreaks.

- **GNNs for Data Analysis** The structured data of the graph contains rich individual connection information. GNNs capture the neural model of graph dependencies through message passing between graph nodes, which can be used to build more robust models [14]. Some work uses GNNs to build the prediction model, for example, Edward et. al. proposed an automated method using GNNs to detect potential links unknown to buyers, and explored how outages can propagate in complex emergency networks [15]. In the field of computational biology, a framework that used graph representation learning for data analysis to obtain global structural features were proposed in [16]; and this technique was also used for clustering scRNA-seq data in [17]. In [18], Yu et. al. used GNNs for traffic flow prediction, which developed an attention mechanism to aggregate information according to adjacency matrix. Although GNNs is often used for location-based data prediction due to its structural characteristics, most of these studies treat the structured information contained as static data, and the edges of the graph can represent dynamic information such as population flow or inter-

regional interaction, which would be very helpful for temporal spatial data analysis.

3. GOPS Framework

The schematic of our framework is shown in Fig. 1. GOPS systems works as follows. First, based on the migration and population, GOPS calculates the edge weights of GNNs. According to them, we updated the input of socio-demographic features considering the influence of their adjacent neighbors. Using the growth of the drugs report quantity and the social group distribution of each county, GOPS contains several machine learning models and deep learning models to make the prediction. Additionally, GOPS allows model ensemble using boosting and bagging, drawing on the wisdom of the masses.

3.1. Opioids Prediction considering Socio-Demography

To simulate the linear growth of the opioids reported quantity, we build the linear regression models to forecast the number of opioid cases in nature growth mode at first. Considering that the homogeneity of a city is often determined by the similarity of population composition, it is intuitive to use the social population distribution of a city as a predictive feature. Because some demographic features are highly overlapped, we need to reduce the dimension of feature vectors, limiting the number of demographic features to a few. To extract the useful features and not hurt the interpretability of the model, we did a rough correlation among these social group elements and coarsely filtered out some features that do have a strong relationship between them. We then calculate the relationship between drug reported amount and social group parameters and assess the importance of social group features according to the Pearson correlation ρ between two weakly dependent features f_1, f_2,

$$\rho_{f_1,f_2} = \frac{Cov(f_1, f_2)}{\sigma_{f_1}\sigma_{f_2}}, \tag{1}$$

where Cov is the covariance operator and $\sigma_{f_1}, \sigma_{f_2}$ are the standard deviations of f_1 and f_2.

3.2. Prediction considering Civil Migration

We plan to use graph neural networks (GNN) to aggregate feature vectors and make our models more robust, based on our notion that opioid increases are caused not just by internal population composition but also by civil migration. Specifically, the aggregated feature vector v_c' of county c can be calculated by

$$v_c' := (1-\alpha)v_c' + \alpha \sum_{i=0}^{n} \mathcal{N}(\frac{m_i^+ + m_i^-}{2p_c})v_i, \tag{2}$$

where n represents the number of neighbor of county, m_i^+ and m_i^- are the migration in and out quantity between county c and i, p_c is the total population of county c, $\mathcal{N}$ is the normalize operator, and α is the updating rate. The hyperparameter α helps to keep the original feature proportionally, thus ensuring their feature vectors are updated on the same scale.

3.3. Correlation with Heart Diseases

Previous studies have shown a strong link between drug abuse and heart disease. In this section, we extend our framework to figure out what types of opioids contribute to the prevalence in heart disease cases. We only focus on the prevalence of common heart diseases like coronary and stroke, due to the fact that the deaths and hospitalization from heart disease can be the result of multiple factors. To solve this issue, we used various correlation calculating methods to find a correlation between the increase of opioids abuse and the prevalence of these heart diseases.

4. Experiments

4.1. Data

Several sets of data from three databases are used in this project.

- **NFLIS Drug Report Dataset** This dataset contains the information of the number of drug reports for various kinds of opioids among counties from 2010 to 2017. In this dataset, we utilize the variables "FIPS-Combined" as the identity for each state. "SubstanceName" represents opioid names. "DrugReports", "TotalDrugReportsCounty", "TotalDrugReportsState" are the quantity values we use for the future forecast.

- **NFLIS Social Group Dataset** This dataset contains the social features for all counties in U.S. There are 152 features contained in this dataset. GEO.id2 is as same as "FIPS-Combined" in NFLIS Drug Report Dataset to identify counties, and left join operators help to merge the tables together.

- **USCensus Population Dataset** USCensus civil population dataset contains the estimated population for each county from 2010 to 2019, it has the geographic area in the first column, following by yearly population.

- **USCensus Civil Migration Dataset** Annual migration data of residents in these five states are included in this dataset. There is Flow from Geography B to Geography A, Counterflow from Geography A to Geography B^2, Net migration from Geography B to Geography A^2, Gross migration between Geography A to

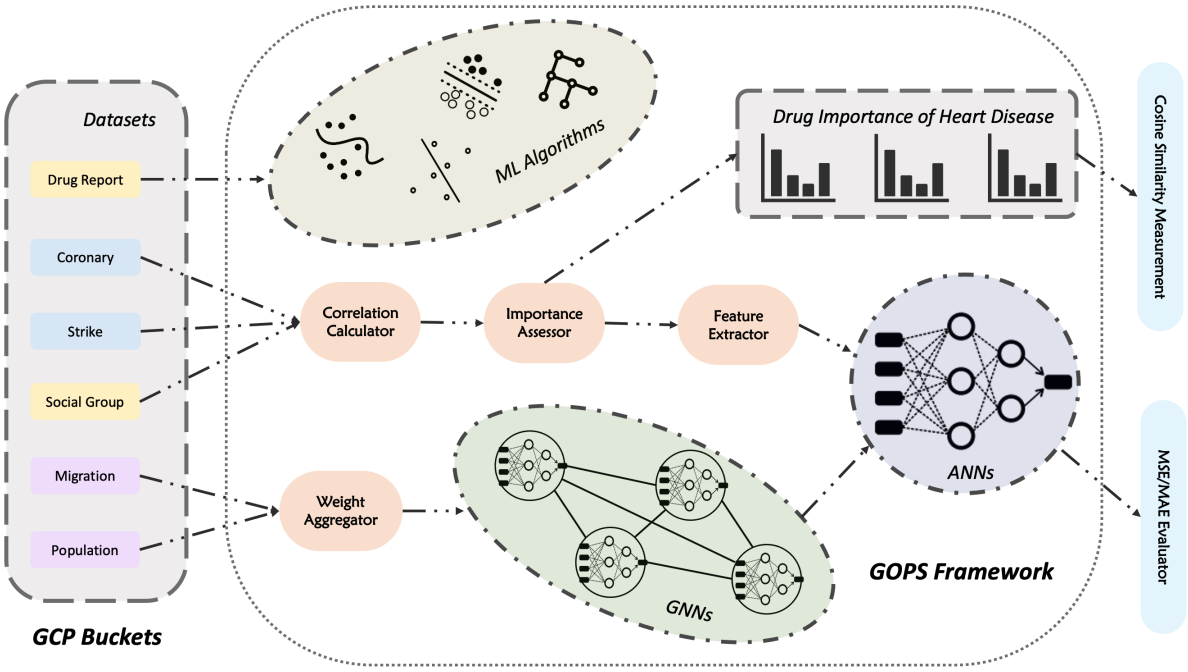

Figure 1: Schematic of our GOPS framework (better to view with color), where datasets with the same color come from the same database.

Geography B^2. In our experiment, we use Flow from Geography B to Geography A, and the "Margin of Error" (MOE) is considered. Based on the data from the USCensus database, we can get the graph of the percentage of inflow and outflow movement of the population of these states between 2010 to 2017.

- **CDC Coronary Heart Disease/Stroke Datasets** These datasets have the prevalence and death rate of people who are affected by heart diseases in each countries. In this paper, we use coronary heart disease and stroke prevalence data in certain counties across the country as examples. For the are certain counties where data points are not available, we set default value 0s for them.

The velocity of our large-volume datasets are coming in fits and starts, sent to us in batches because the census data are not real-time. All of our datasets are in table format, either in csv or txt, but are very different representability of contents (columns, values-rate or quantity).

4.2. Tools and Platform

We conduct the experiments on Google Cloud Platform (GCP) and use Google Compute Engine including Dataproc as our primary service because of its powerful feature of easily manage, process, and visualize the data, as well as the salable and reliable services.

In our experiments, we create clusters with default configuration in a single-region setting. We built our system using tools such as Clusters, BigQuery (to filter and choose important data), and single-region cloud storage (to store enormous volumes of data). To fully utilize the power of clusters, we store our data into GCP buckets and take advantage of BigQuery putting those datasets into the tables. We use BigQuery with postgresql queries to extract useful information, to conduct the data cleaning and pre-processing, and to join the datasets in different database. We build correlation calculators, weight aggregators, feature extractors, ANNs and GNNs individually on different clusters for their different configurations.

Base on those settings, we pick PySpark and Scikit-Learn as base tools to build our models. Their well-defined interface provides us a lot of advanced and decent functions that could speed up our project. Other than that, we also introduce Tensorflow and Keras to build the ANNs and GNNs. After that, we use seaborn and matplotlib to visualize our project results.

4.3. Metrics

We use Minimum Square Error (MSE) and Minimum Absolute Error (MAE) as the metrics for evaluating our regression models according to the Eq. 3. In our experiment, data in 2010-2015 are used as training set and data in 2016-2017 are used as validation set. The models with smaller

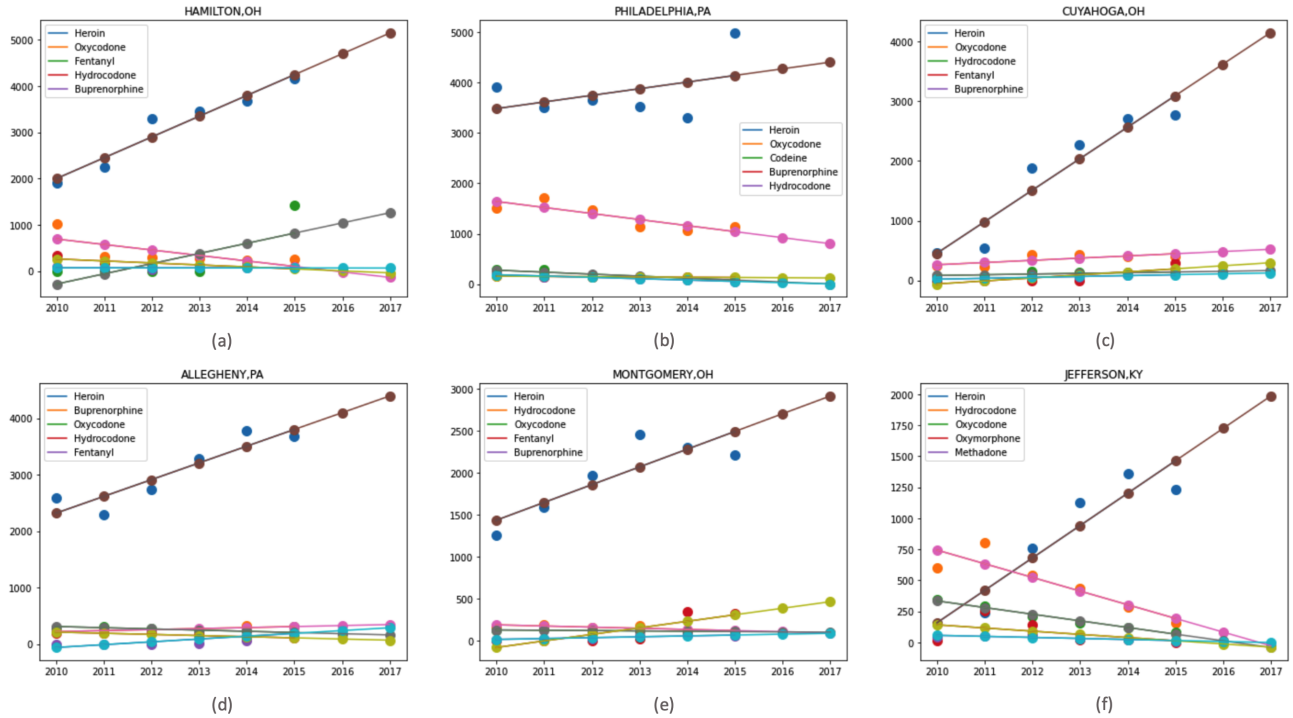

Figure 2: Linear growth prediction results for 6 counties, i.e., (a) Hamilton, OH, (b) Philadelphia, PA, (c) Cuyahoga, OH, (d) Allegheny, PA, (e) Montgomery, OH, (f) Jefferson, KY. Scatters are the observation points and line plots are the linear growth models.

MSE and MAE are better trained and vice versa.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - f(\boldsymbol{x}_i))^2, \tag{3}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{y}_i - f(\boldsymbol{x}_i)),$$

where n is the number of data points, $\boldsymbol{y}_i$ are the ground truths in the validation set, $f(\boldsymbol{x}_i))^2$ are the predicted values of $\boldsymbol{x}_i$ by our model f.

We employ cosine similarity measurement for evaluating the correlation analysis using different methods according to the Eq. 4.

$$\cos(\boldsymbol{x},\boldsymbol{y}) = \frac{\boldsymbol{x}\cdot\boldsymbol{y}}{||\boldsymbol{x}|| \times ||\boldsymbol{y}||} \tag{4}$$

The correlation method with biggest similarity, i.e., the smallest cosine value, has the greatest robustness and therefore should be chosen as our calculation method. In our experiment, the two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ are the coefficients calculated by different correlation methods in training and all datasets.

4.4. Prediction of Linear Growth

We built linear regression models to predict the reported opioid quantities as baselines and used MSE and MAE as metrics. We developed a basic linear regression model that only comprised internal components, with opioid reported quantity data from 2010 to 2015 serving as the training set and data from 2016 to 2017 serving as the validation set, by setting an adaptive learning rate from 0.05 to 0.01 with stochastic gradient descent.

In this experiment, we examined 464 counties from 5 states and the results of selected 6 counties are shown in Fig. 2. As we can see from the graphs, some kinds of opioid reports are increasing non-linearly in some regions, indicating that we may need more sophisticated machine learning models to make the prediction.

4.5. Prediction considering Socio-Demography

To take the socio-demogrphy into consideration, we employed several machine learning models from Scikit-Learn and PySpark interfaces, like support vector regression, ridge regression, naive Bayes regression, CART decision tree regression, etc. We calculated the correlation between 152 features and drug reported quantity growth, and manually selected the social group population ratios of 10 independent (the absolute values of correlation coefficients are no larger than 0.4) and important features as input vectors according to the correlation between features. The correlation of relatively independent features are shown in Fig. 7 in

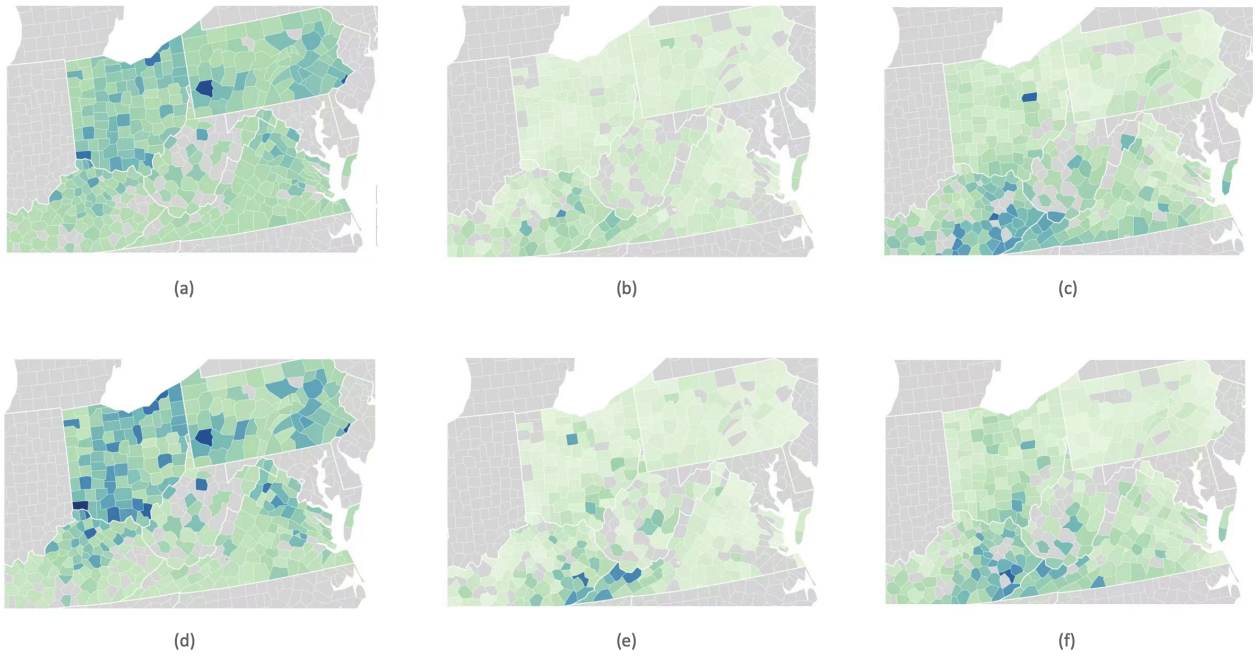

Figure 3: The heat-maps of the percentage of the predicted reported quantity rate to the county population of 3 opioids of all counties in 5 U.S. states by fully connected networks and enhanced GNN models: (a) heroin prediction by fully connected networks; (b) buprenorphine prediction by fully connected networks; (c) oxycodone prediction by fully connected networks; (d) heroin prediction by GNNs; (e) buprenorphine prediction by GNNs; (f) oxycodone prediction by GNNs. The scale of color bar is from 0% to 0.35% (the total reported quantities do not exceed 35 per 10,000 of the total population). The counties with no colors are missing data.

County	Linear Regression			Support Vector Regression			Ridge Regression			Fully Connected Networks			Ground Truth		
	A	B	C	A	B	C	A	B	C	A	B	C	A	B	C
Hamilton, OH	5014	0	0	1058	861	1023	2078	32	341	2758	77	287	4525	81	221
Philadelphia, PA	4102	12	1033	2012	1647	1909	3951	17	809	4360	128	773	5075	136	865
Cuyahoga, OH	4024	123	366	1688	1362	1444	2988	44	440	2855	114	398	2735	106	404
Allegheny, PA	4233	28	251	1613	1336	1400	2423	58	144	3612	221	177	3400	293	183
Montgomery, OH	2964	263	0	705	559	562	1788	47	214	1045	91	186	1504	123	143
Jefferson, KY	1998	0	3	977	768	875	2736	108	336	1002	54	221	1000	17	69
MSE	62561	775	2563	83630	40149	33699	59472	731	1463	14181	128	386	0	0	0
MAE	79	27	43	86	91	88	83	13	21	34	6	11	0	0	0

Table 1: The prediction results and ground truths (values in parentheses) of 2 selected machine learning models (support vector regression and ridge regression) and fully connected networks with 6 layers ($11 \times 16 \times 26 \times 10 \times 4 \times 1$) of 6 counties in 2017. A, B, C represent heroin, buprenorphine, and oxycodone respectively. The last two lines of the table shows the MSE and MAE of each model.

the Appendix. By feature extracting, we found widowed males over 15 years old, separated females over 15 years old, people over 25 years old with less than 9th grade, children under 18 with a disability, and etc., are more likely to be addicted to opioids.

Among these models, the performance of ridge regression and support vector machine regression perform the best. Specifically, we set the conjugate gradient solver as ridge regression solver for large-scale data and the regularization strength as 1.0; for support vector regression, we set

radial basis function as kernel and 0.1 as the kernel coefficient. For the deep learning models, after fine-tuning the network parameters, we used fully connected neural networks with 6 layers ($11 \times 16 \times 26 \times 10 \times 4 \times 1$ neurons), where sigmoid function is used as activation function in the last layer. We used Adam optimizer to train 100 epochs with batch size 10, and adopt binary cross entropy as our loss function.

Their prediction results of 3 opioids in 6 example counties are shown in Tab. 1. In addition, to get a whole picture of the prediction in all counties of these 5 states, the heat-maps of 3 opioids predictions are shown in Fig. 3 (a-c).

From Tab. 1., we can see that all of these machine learning models outperform the simple linear regression models, where the deep neural network with fine-tuned parameters can obtain the best performance with smallest MSE and MAE, i.e., 1085, 12. Through the heat-maps, we predict the use of drug following counties could surge in the future, leading to an opioid crisis in that region: heroin in Allegheny county, PA, Cuyahoga county, OH, Pickaway county, OH; buprenorphine in Lee county, KY; oxycodone in Holmes county, OH, Lee county, KY.

4.6. Prediction considering Civil Migration

Our GNN model takes into account the transportation of goods (drugs) and the changes of composition of the population due to the migration of residents. For this model, we used one-time aggregation with 0.01 learning rate. The normalized aggregation weights of the 5 states above are shown in Tab. 2.

	KY	OH	PA	VA	WV
KY	0.00	0.61	0.07	0.14	0.18
OH	0.37	0.00	0.17	0.10	0.36
PA	0.07	0.28	0.00	0.31	0.34
VA	0.13	0.15	0.27	0.00	0.45
WV	0.11	0.37	0.21	0.31	0.00

Table 2: The normalized aggregation weights for 5 U.S. states calculated by migration rate, i.e., the percentage of the average of in and out flow to its population.

After fusing the feature vectors, we fed the updated features into the same deep learning models in the above section and get the results in Tab. 3. The heat-maps of 3 opioids predictions are shown in Fig. 3 (d-f). Similarly, we predict that the abuse of opioids could increase rapidly in the near future in the following hotspots: heroin in Hamilton county, PA, Allegheny county, PA, Cuyahoga county, PA; buprenorphine in Perry county, KY, Buchanan county, VA; oxycodone in Perry county, KY, Knott, KY, Buchanan county, VA.

County	GNN			Ground Truth		
	A	B	C	A	B	C
Hamilton, OH	5120	74	861	4525	81	221
Philadelphia, PA	4698	135	873	5075	136	865
Cuyahoga, OH	3571	168	732	2735	106	404
Allegheny, PA	3156	321	212	3400	293	183
Montgomery, OH	2689	134	161	1504	123	143
Jefferson, KY	1008	47	216	1000	17	69
MSE	13657	131	380	0	0	0
MAE	30	7	10	0	0	0

Table 3: The prediction results and ground truths (values in parentheses) of GNN models with 0.01 aggragation rate of 6 counties in 2017, where the local networks have the same structure as the ones above. A, B, C represent heroin, buprenorphine, and oxycodone respectively. The last two lines of the table shows the MSE and MAE of the model.

From Tab. 3, we can see GNNs slightly enhance the model and reduce the MSE 3% and MAE 13% for some opioids. This method has an even more important advantage, which is to improve the robustness of the system by reducing the dependence of local networks on specific features. This will be covered in more detail in the sensitivity analysis section.

4.7. Correlation with Heart Diseases

To find out the potential impact of different opioids on heart disease like coronary and stroke, we used combined prevalence to calculate the correlation coefficient and used its difference with majority voting of each county to estimate the error. Fig. 4 shows the correlation between various kinds of opioids with heart diseases using Pearson correlation method.

As we can see in Fig. 4 (a), heroin and morphine are positively related to heart disease's prevalence in 4 states. While, other opioids do not demonstrate clear relationship across these states, showing that heroine and morphine abuse may lead to the increase of coronary. In addition, according to Fig. 4 (b), heroin is positively related to heart disease's prevalence in all states except for Virginia, and morphine is positively related to heart disease's prevalence in all states except for Philadelphia. Consequently, we can speculate that heroin and morphine abuse may lead to cardiovascular disease.

For different correlation methods, we evaluated their performance by employing cosine similarity measurement between the training sets and the whole sets. The evaluation results are shown in Tab. 4. We can see that choosing Pearson correlation method is reasonable in terms of the consis-

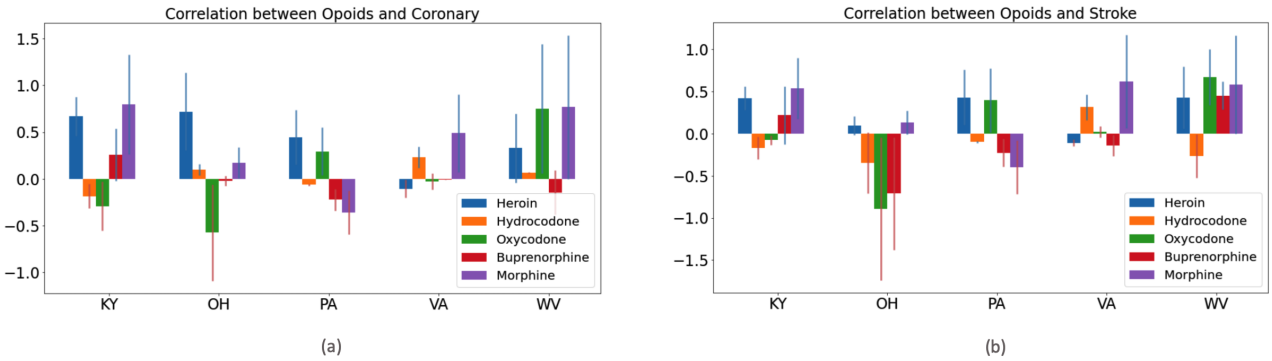

Figure 4: (a) Pearson correlation between 5 different kinds of opioids and coronary increase for 5 U.S. states; (b) Pearson correlation between 5 different kinds of opioids and stroke increase for 5 U.S. states. Blue error bars mean the over-estimations and red ones mean the under-estimations.

tency of the different parts of the dataset, since it has the maximum similarity for these two heart disease.

	Pearson	**Kendall**	**Spearman**
Coronary	87.08%	83.97%	85.44%
Stroke	85.79%	85.21%	86.67%
Mean	86.44%	84.59%	86.06%

Table 4: The similarity of between the training sets and the whole sets for 3 different correlation methods, i.e., Pearson, Kendall, Spearman.

4.8. Sensitivity Analysis

To test the robustness of our GOPS framework, we tested how the models behave under different hyperparameter fluctuations. Specifically, we divided the parameters of GOPS into three categories, i.e., graph-level, node-level, and feature-level parameters and conduct the sensitivity analysis that explores the stability of GOPS from three perspectives of aggregation strength, network structure, and feature extraction. The parameter fluctuation analysis from these three levels is shown in Fig. 5.

- **Graph-Level** In the GNN, the learning rate determines "how much a model listens to its neighbor", thus it can change a lot for model performance. We tuned the learning rate of GOPS from 0.05 to 0.2 and the evaluation results of our models in predicting oxycodone are show in Tab. 5 lr - α column. Similarly, the number of aggregation times also affect the performance of the model, but different from the learning rates in one time that directly consider of neighbor influence, more iterations contain more evolutionary information, thus could be able to represent the real growth trend in the process of natural population migration process.

  The evaluation results of our models are show in Tab. 5 # agg_iter column. From the results, we found that, for some kinds of opioids like oxycodone and fentanyl, considering the neighboring effects could help to improve the performance, but adopting too much might further destabilize the model, i.e., when $\alpha > 0.10$ in oxycodone and $\alpha > 0.05$ in fentanyl. And generally, the performance of the models do not depend lots on this parameter, in other words, it is not sensitive. As for the aggregation iterations, increasing the value of this parameter will devastate the models in a large scale, it is probably because the learning rate should be decrease accordingly when increasing this parameter.

- **Node-Level** In GNN, each node is a artificial neural network or machine learning model, we focus on the network structures here, i.e., the number of neurons in each layer. Knowing the sensitivity of number of neurons could help us to fine tune the model structure, hence get better prediction results. Here we tuned the number of neurons of the first three hidden layers from 10-20, 20-30, and 5-15 respectively, and selected some of the evaluation results of predicting oxycodone in Tab. 5 # neuron column.

  By observing the low-lying areas in Fig 5 (c) and (d), we found that the number of neurons in the layers of most models with better performance was concentrated at $(11 \times 16 \times 22 \times 10 \times 4 \times 1)$, and the model performance could not be continuously improved by increasing the number of neurons. In this experiment, the model performance is insensitive to the number of neurons in the fully connected networks.

- **Feature-Level** Before making the prediction, we need

Metric	lr - α					# agg_iter			# neuron					# feature			
	0.00	**0.05**	**0.10**	**0.15**	**0.20**	**1**	**2**	**3**	**a**	**b**	**c**	**d**	**e**	**5**	**10**	**20**	**38**
MSE	386	382	380	397	416	380	396	475	1611	1517	1598	1588	1892	2211	1543	2562	2752
MAE	11	11	10	13	16	10	13	18	21	19	18	21	23	22	20	21	23

Table 5: The evaluation results of predicting oxycodone using different hyperparameter settings in terms of MSE and MAE. lr - α the learning rate in GNNs aggregating process; # agg_iter is the GNNs aggregation times; # neuron is the number of neurons in the node networks; # feature is the number of extracted features. In # neuron, a, b, c, d, e represent 6-layer networks with $(11 \times 10 \times 20 \times 5 \times 4 \times 1)$, $(11 \times 10 \times 30 \times 15 \times 4 \times 1)$, $(11 \times 15 \times 25 \times 10 \times 4 \times 1)$, $(11 \times 20 \times 20 \times 5 \times 4 \times 1)$, $(11 \times 20 \times 30 \times 10 \times 4 \times 1)$ neurons respectively.

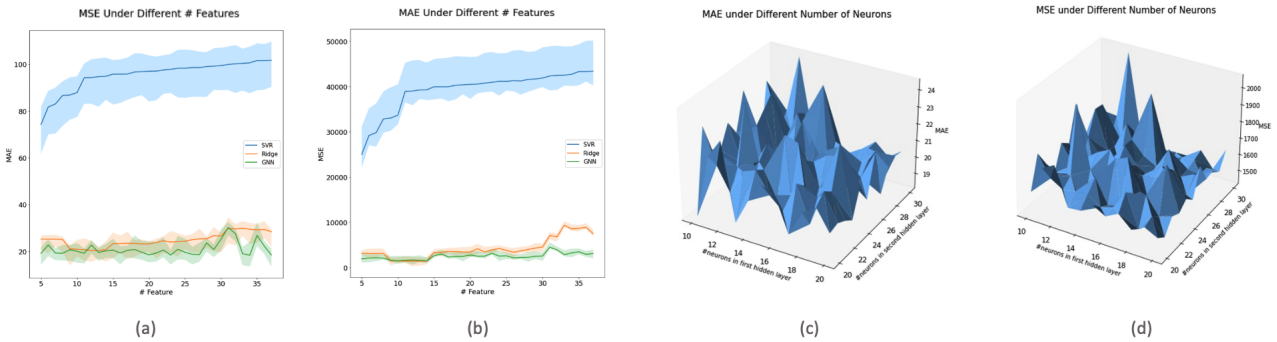

Figure 5: Sensitivity analysis of hyperparameters from graph-level, node-level, and feature-level: (a) The MSE of various machine learning models under different number of features (5-38), the error bands are the fluctuations of learning rates (0.05-0.40); (b) The MAE of various machine learning models under different number of features (5-38), the error bands are the fluctuations of learning rates (0.05-0.40); (c) MSE under different number of neurons of first two hidden layers (10-20, 20-30); (d) MAE under different number of neurons of first two hidden layers (10-20, 20-30).

to extract important features to reduce the dimension, whose number could significantly affects the model performance. We set the extracted feature number parameter from 5-38 to explore "how many features are best used to make predictions". Although this approach looks somewhat similar to PCA analysis, however, to preserve interpretability, we customized the way we filter features, thus introducing this extra hyperparameter. The evaluation results in predicting oxycodone are shown in Tab. 5 # feature column.

From this column, we can see that the number of features do not contribute to the models and even do little harm to them after passing a threshold, e.g., 10 for oxycodone. This reflects the necessity and effectiveness of reducing the dimension of the input data.

The analysis indicates that, the models are not sensitive to most of the hyperparameters except the aggregation times. This shows the effectiveness of our manual feature extraction, and by properly setting up, GOPS framework is robust enough to handle opioid analysis tasks.

5. Conclusion

In this paper, we propose a big data analysis framework GOPS to make the prediction for a recent hot topic opioid crisis. GOPS uses several machine learning model and is deployed in GCP in well manner. Taking the geographical influence into consideration, GOPS make the opioid prediction using the socio-demographic features. Based on preset thresholds and hyperparameters, GOPS can predict with relative accuracy the locations of future opioid outbreaks. In addition, the sensitivity analysis for GOPS indicates that GOPS is robust and not sensitive to most hyperparameter changes, which reflects the feasibility of its future migration to industrial systems. This framework has high commercial and medical value: it can help local governments increase the drug supervision of certain fast-growing and highly induced drugs, and the drug supervision of certain special social groups; at the same time, it can guide hospitals to local possibilities to prevent the outbreak of heart disease.

Though, GOPS still has room for improvement. In the future, we plan to extend it to make it adaptable for continuous variables (e.g., age, income, etc.), matrix inputs, as well as the mixture. In addition, we are looking to take

more diseases into consideration from a more holistic perspective. In the end, we will try to unify drugs in the future to identify the addictive ingredients and harmful ingredients by comparing their growth trend and influence.

References

[1] Centers for Disease Control and Prevention (CDC). Cdc grand rounds: prescription drug overdoses - a u.s. epidemic. 1

[2] H.B. Hafeiz. Socio-demographic correlates and pattern of drug abuse in eastern saudi arabia. *Drug and Alcohol Dependence*, 38(3):255–259, 1995. 1

[3] Goldfrank L Chiang W. The medical complications of drug abuse. *Pain Medicine*, 152(2):83–88, 1990. 1

[4] Stefan Kertesz Yulia Khodneva, Paul Muntner. Prescription opioid use and risk of coronary heart disease, stroke, and cardiovascular death among adults from a prospective cohort (regards study). *Pain Medicine*, 17(12):444–455, 2016. 1

[5] Jonathan H Chen and Steven M Asch. Machine learning and prediction in medicine - beyond the peak of inflated expectations., 2017. 1

[6] Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of Biomedical Informatics*, 69:218–229, 2017. 1

[7] Patrizia Villari Sebastiano Mercadante, Patrizia Ferrera. Frequency, indications, outcomes, and predictive factors of opioid switching in an acute palliative care unit. *Journal of Pain and Symptom Management*, 37(4):632–641, 2009. 2

[8] Kenneth S Kendler, Kristina Sundquist, Henrik Ohlsson, Karolina PalmÚr, Hermine Maes, Marilyn A Winkleby, and Jan Sundquist. Genetic and familial environmental influences on the risk for drug abuse: a national swedish adoption study. *Archives of general psychiatry*, 69(7):690–697, 2012. 2

[9] Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, and Shanfeng Zhu. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in Bioinformatics*, 15(5):734–747, 08 2013. 2

[10] ShanghaiTech University. Team 1906204: Analysis of the opioid crisis and strategies. 2

[11] University of Colorado Boulder. Team 1900577: The gravity of the opioid crisis. 2

[12] University of Colorado Boulder. Team 1901213: Take me home: Preventing journeys down the opioid addiction road. 2

[13] University of Colorado Boulder. Team 1901679: Random walks and rehab: Analyzing the spread of the opioid crisis. 2

[14] Shengding Hu Jie Zhou, Ganqu Cui. Graph neural networks: A review of methods and applications. *AI Open*, 1:57–81, 2020. 2

[15] Edward Elson Kosasih and Alexandra Brintrup. A machine learning approach for predicting hidden links in supply chain with graph neural networks. *International Journal of Production Research*, 0(0):1–14, 2021. 2

[16] Nicolas Swenson, Aditi S Krishnapriyan, Aydin Buluc, Dmitriy Morozov, and Katherine Yelick. Persgnn: Applying topological data analysis and geometric deep learning to structure-based protein function prediction. *arXiv preprint arXiv:2010.16027*, 2020. 2

[17] Madalina Ciortan and Matthieu Defrance. GNN-based embedding for clustering scRNA-seq data. *Bioinformatics*, 11 2021. btab787. 2

[18] Xiaoyang Wang, Yao Ma, Yiqi Wang, Wei Jin, Xin Wang, Jiliang Tang, Caiyan Jia, and Jian Yu. Traffic flow prediction via spatial temporal graph neural network. In *Proceedings of The Web Conference 2020*, WWW '20, page 1082–1092, New York, NY, USA, 2020. Association for Computing Machinery. 2

Appendix

Statics of Total Drug Reported Quantity

In this paper, we focus on the following opioids in the Tab. 6.

Cyclopropyl fentanyl	Dihydrocodeine	Thebaine	Opiates	U-47700
Benzylfentanyl	Fluorobutyryl fentanyl	Cyclopropyl/Crotonyl Fentanyl	4-Methylfentanyl	Furanyl fentanyl
p-methoxybutyryl fentanyl	o-Fluorofentanyl	Acetyldihydrocodeine	Methadone	p-Fluorofentanyl
Tetrahydrofuran fentanyl	3-Methylfentanyl	Hydrocodone	Opium	Crotonyl fentanyl
Buprenorphine	Oxycodone	Fluorofentanyl	4-Fluoroisobutyryl fentanyl	Phenyl fentanyl
p-Fluorobutyryl fentanyl	Fentanyl	Acryl fentanyl	Meperidine	Desmethylprodine
trans-3-Methylfentanyl	Propoxyphene Butorphanol	Methorphan	U-48800	Tramadol
Pentazocine	Acetyl fentanyl	Acetylcodeine	Pethidine	Codeine
Dextropropoxyphene	Carfentanil	cis-3-methylfentanyl	U-51754	Fluoroisobutyryl fentanyl
Methoxyacetyl fentanyl	Butyryl fentanyl	Heroin	Valeryl fentanyl	Mitragynine
Hydromorphone	ANPP	Morphine	U-49900	Oxymorphone

Table 6: The opioids that we studied in this paper.

The statistics of the NFLIS dataset is in Fig. 6, where (a) shows the percentage of some selected opioids (report quantity larger than 10 in total), and (b) shows the map whose color are determined by the number drug reports. We can see that the heroin, oxycodone, hydrocodone, fentanyl, and buprenorphine make up the most percentage of the opioid reports.

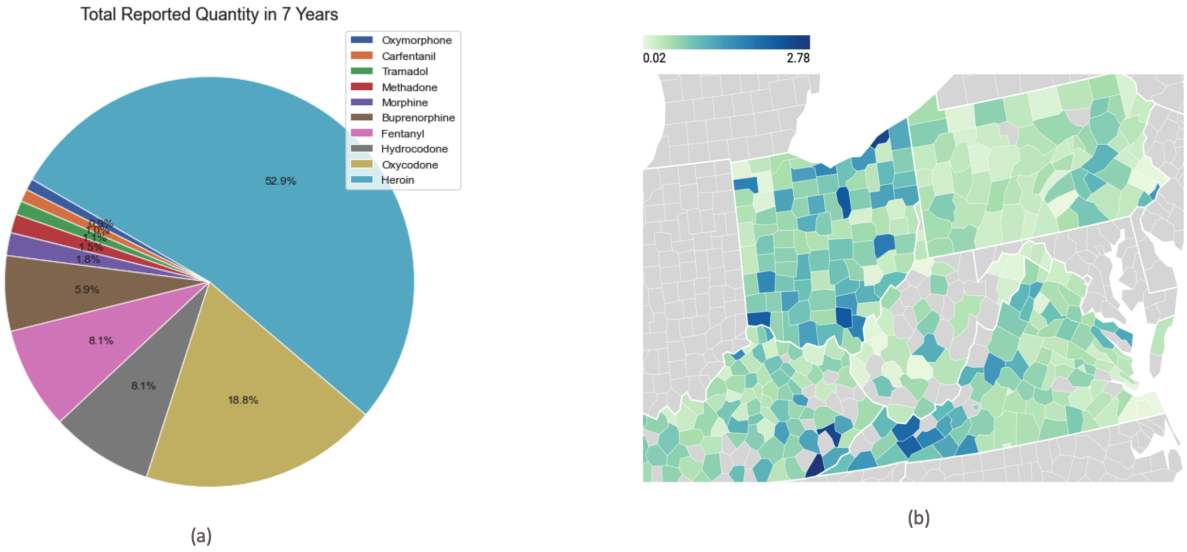

Figure 6: The statistics of the NFLIS dataset: (a) The pie plot of the total drug reported quantity in 2010-2017; (b) The percentage of the total drug reported quantity to the local population.

Manually Dimension Reduction Process

When extracting data, we first rough filtered the social group features by setting 0.4 as threshold for their coefficients and obtained 38 relatively independent features. Their correlation are shown in Fig. 7.

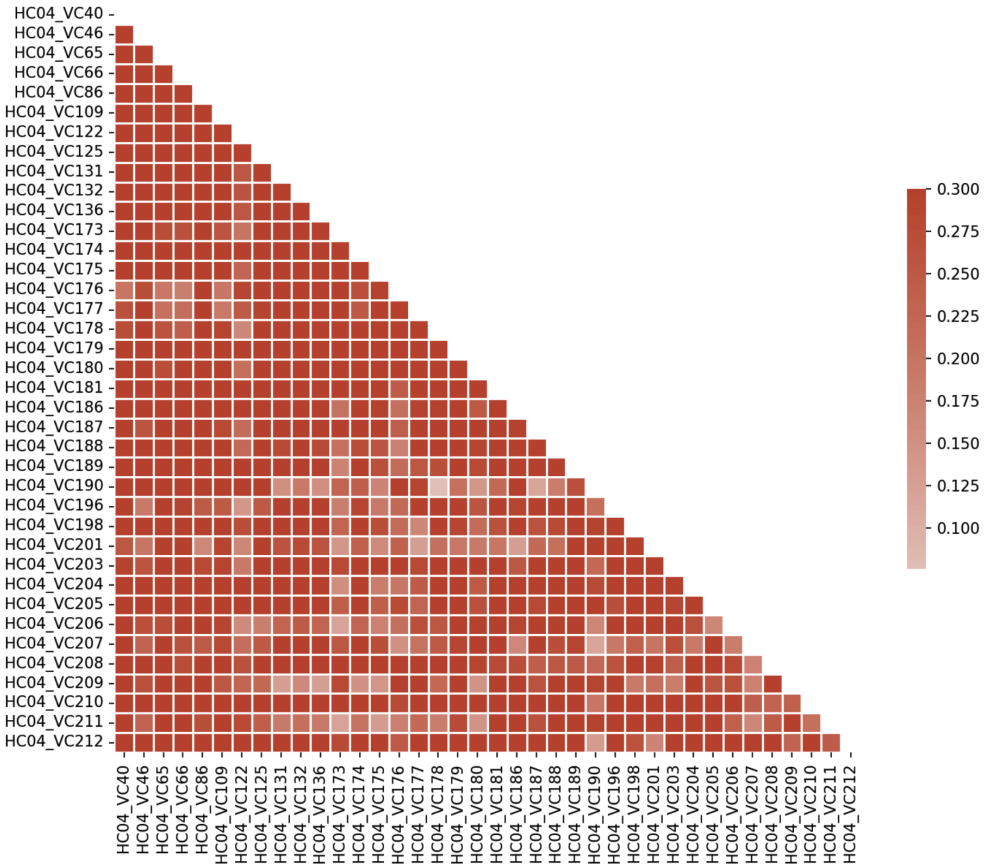

Figure 7: Correlation matrix between 38 independent social group features (whose coefficient no larger than 0.4), according data from 2010 to 2015.

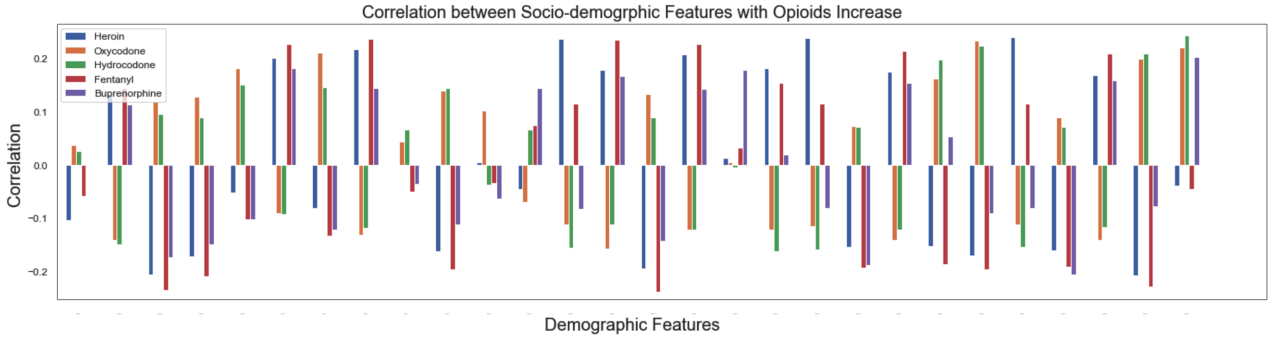

Figure 8: The correlation between 38 relatively independent features with the increase of different opioids.

Then according to their correlation with opioid reports as shown in Fig. 8, we select the top-K important features. The socio-demographic codes and their names are shown in Tab. 7. In this way, we reduce the dimension of input data, and keep the intepretability of the extracted features at the meantime.

Code	Socio-Demographic Features
HC04_VC40	MARITAL STATUS - Males 15 years and over - Widowed
HC04_VC46	MARITAL STATUS - Females 15 years and over - Separated
HC04_VC65	GRANDPARENTS - Number of grandparents living with own grandchildren under 18 years - Years responsible for grandchildren - Less than 1 year
HC04_VC66	GRANDPARENTS - Number of grandparents living with own grandchildren under 18 years - Years responsible for grandchildren - 1 or 2 years
HC04_VC86	EDUCATIONAL ATTAINMENT - Population 25 years and over - Less than 9th grade
HC04_VC109	DISABILITY STATUS OF THE CIVILIAN NONINSTITUTIONALIZED POPULATION - Under 18 years - With a disability
HC04_VC122	RESIDENCE 1 YEAR AGO - Population 1 year and over - Different house in the U.S. - Same county
HC04_VC125	RESIDENCE 1 YEAR AGO - Population 1 year and over - Different house in the U.S. - Different county - Different state
HC04_VC131	PLACE OF BIRTH - Total population - Native
HC04_VC132	PLACE OF BIRTH - Total population - Native - Born in United States
HC04_VC136	PLACE OF BIRTH - Total population - Foreign born
HC04_VC173	LANGUAGE SPOKEN AT HOME - Population 5 years and over - Language other than English - Speak English less than ”very well”
HC04_VC174	LANGUAGE SPOKEN AT HOME - Population 5 years and over - Spanish
HC04_VC175	LANGUAGE SPOKEN AT HOME - Population 5 years and over - Spanish - Speak English less than ”very well”
HC04_VC176	LANGUAGE SPOKEN AT HOME - Population 5 years and over - Other Indo-European languages
HC04_VC177	LANGUAGE SPOKEN AT HOME - Population 5 years and over - Other Indo-European languages - Speak English less than ”very well”
HC04_VC178	LANGUAGE SPOKEN AT HOME - Population 5 years and over - Asian and Pacific Islander languages
HC04_VC179	LANGUAGE SPOKEN AT HOME - Population 5 years and over - Asian and Pacific Islander languages - Speak English less than ”very well”
HC04_VC180	LANGUAGE SPOKEN AT HOME - Population 5 years and over - Other languages
HC04_VC181	LANGUAGE SPOKEN AT HOME - Population 5 years and over - Other languages - Speak English less than ”very well”
HC04_VC186	ANCESTRY - Total population - American
HC04_VC187	ANCESTRY - Total population - Arab
HC04_VC188	ANCESTRY - Total population - Czech
HC04_VC189	ANCESTRY - Total population - Danish
HC04_VC190	ANCESTRY - Total population - Dutch
HC04_VC196	ANCESTRY - Total population - Hungarian
HC04_VC198	ANCESTRY - Total population - Italian
HC04_VC201	ANCESTRY - Total population - Polish
HC04_VC203	ANCESTRY - Total population - Russian
HC04_VC204	ANCESTRY - Total population - Scotch-Irish
HC04_VC205	ANCESTRY - Total population - Scottish
HC04_VC206	ANCESTRY - Total population - Slovak
HC04_VC207	ANCESTRY - Total population - Subsaharan African
HC04_VC208	ANCESTRY - Total population - Swedish
HC04_VC209	ANCESTRY - Total population - Swiss
HC04_VC210	ANCESTRY - Total population - Ukrainian
HC04_VC211	ANCESTRY - Total population - Welsh
HC04_VC212	ANCESTRY - Total population - West Indian (excluding Hispanic origin groups)

Table 7: The code of 38 relatively independent socio-demographic features and their meanings.