A Novel Big Data Analytics Framework to Predict the Risk of Opioid Dependency

Md Mahmudul Hasan[1], Md. Noor-E-Alam[1*], Mehul Rakeshkumar Patel[1], Alicia Sasser Modestino[2,3], Gary Young[4,5,6]

[1] Mechanical and Industrial Engineering, Northeastern University, Boston, Massachusetts, United States of America,

[2]Public Policy and Urban Affairs and Economics, Northeastern University, Boston, Massachusetts, United States of America

[3]Dukakis Center for Urban and Regional Policy, Northeastern University, Boston, Massachusetts, United States of America

[4] D'Amore-McKim School of Business, Northeastern University, Boston, Massachusetts, United States of America

[5] Center for Health Policy and Healthcare Research, Northeastern University, Boston, Massachusetts, United States of America

[6] Bouvé College of Health Sciences, Northeastern University, Boston, Massachusetts, United States of America

Corresponding author: Md. Noor-E-Alam, Northeastern University, School of Engineering, 360 Huntington Avenue, Boston MA 02115, email: mnalam@neu.edu, Tel.: +1-617-373-2275

Abstract

Addiction and overdose related to prescription opioids have reached an epidemic level in the U.S., creating an unprecedented national crisis. This has been exacerbated partly due to the lack of tools for physicians to help predict whether or not a patient will develop dependency on opioids. Prior research lacks the investigation of how machine learning can be applied to a big-data platform to ensure an informed and judicious prescribing of opioids. In this study, we explore the Massachusetts All Payer Claim Data (MA APCD), a de-identified healthcare claim dataset, and propose a novel framework to examine how naïve users develop dependency on opioids. We perform several feature engineering techniques to identify the influential demographic and clinical features associated with opioid dependency from a class imbalanced analytic sample. We then use and compare the predictive power of four well-known machine learning algorithms: logistic regression, random forest, decision tree, and gradient boosting, to predict the risk of such dependency. Results showed that the random forest model outperforms the other three algorithms while determining the features, some of which are consistent with prior clinical findings. We anticipate that this research has the potential for healthcare practitioners to improve the current prescribing practice of opioids, thereby curbing the increasing rate of opioid addiction.

1. Introduction

The opioid addiction epidemic—a recently declared national public health emergency—is the cause of an average 115 deaths in the US every day; 44% of which are attributable to the overdose of legally obtained prescription opioids [1]. Opioids were involved in 42,249 deaths in 2016, and opioid overdose deaths were five times higher in 2016 than 1999, the largest proliferation in overdose related death ever recognized in this country's history [2]. This is a statistic that is difficult to fully grasp—shocking in its scope and overwhelming in the utter complexity of trying to curtail the overflow of legal and illegal drugs that are drowning the whole nation in a ravaging crisis. Consequently, in October 2017 the opioid epidemic was declared a national public health emergency. Although strong evidence for the long-term benefits of opioid therapy for chronic pain management is still lacking [3], 249 million prescriptions were written by physicians in 2013, which is enough for every person in the U.S. to have a bottle of pills in their medicine cabinet [4].

Lack of stringent restrictions on the flow of prescription opioids has enabled people to become increasingly dependent on opioids even without an initial prescription, forcing the whole nation to embrace a pressing epidemic of addiction and overdose. At an economic level, opioid addiction interferes with worker output and productivity. Annually, national opioid abuse-related costs are more than $700 billion

[5]. At a medical level, opioid addiction puts users at risk of other dangerous diseases and conditions including HIV, hepatitis, cirrhosis, and cognitive decline [5].

A key element of the problem is that physicians lack tools for evaluating whether a patient is likely to misuse and ultimately become addicted to opioid medications prescribed for pain. Currently, when prescribing opioids, physicians often use risk assessment tests based on a brief risk interview. Prior study has showed that the Opioid Risk Tool (ORT) and Screener and Opioid Assessment for Patients with Pain-Revised (SOAPP-R), which are used nationwide, inadequately predict which patients will be misusing their opioid medications [6]. Thus, opioids are being prescribed at an alarming rate and current prescribing methods are inefficient, a duo which results in a large amount of opioid misuse and death. Opioid-related addiction and fatalities could be prevented if better opioid prescribing methods could be developed.

Identification of individuals prone to develop opioid dependency is of paramount importance, given the substantial recent increases in prescription rates and consequent increases in opioid related morbidity and mortality. However, there has been little in-depth research on predicting the risk of opioid dependency beforehand so that physicians might prescribe them in an informed way. Previous research has been confined to small samples and narrow clinical and socio-demographic risk factors [7, 8]. In addition, commonly used multi-variate logistic regression models suffer from low predictive accuracy when applied on imbalanced datasets (i.e. the number of patients dependent on opioids is significantly smaller than the non-dependent patients), thus limiting the reliability and applicability of the risk prediction methods to detect opioid dependency [7, 9-12]. This limitation is largely driven by the absence of computationally efficient and robust predictive algorithms that can handle large-scale and imbalanced datasets. Hence, there exists a critical need to comprehensively identify patient-level factors that collectively escalate the risk of becoming dependent on opioids and ensure judicious opioid-prescribing practices.

This study intends to provide the knowledge and tools necessary for improving opioid-prescribing practice s and thus help curb the country's rapidly increasing rate of opioid addiction. The essential idea is to harness the potential of big-data analytics to identify the patient-level factors associated with opioid dependency, and subsequently develop a model to predict the likelihood of opioid addiction. The motivation behind this is that research leveraging advanced big-data analytics to precisely and proactively predict the risk of opioid dependency is still in its infancy. Yet, cutting edge data-mining and machine learning techniques have been demonstrated to perform with higher accuracy in predicting risk scores in a variety of healthcare applications. We anticipate that big-data analytics can potentially transform information into knowledge about how a naïve opioid patient progresses towards opioid dependency, which can support the development of future interventional strategies to prevent subsequent drug dependency. Our rationale for this research is that the ***Massachusetts All Payer Claims Datasets (MA APCD)*** [13], *a*

large-scale de-identified claims database entailing pharmacy claims, medical claims, and member eligibility data contains a wealth of information that can be exploited to investigate the pattern of patients' long-term dependency on opioids and associated risk factors. The results of our investigation will help physicians predict the risk of patient dependency on opioids at a later stage.

The key contributions of this of this study are:

i. We develop a novel framework that can analyze large scale healthcare administrative datasets in an effort to investigate patients' long-term opioid usage pattern. Leveraging this, the proposed framework reveals the trajectory of how a naïve opioid patient later develops dependency on those drugs, and also determines the influential features associated with opioid dependency. It is true that currently there is a paucity of structured datasets populated with influential risk factors of opioid dependency—a fact that has made the development of a reliable predictive model a challenging task. However, with our study we show that the use of predictive analytics to inform judicious opioid prescribing practice is feasible with administrative claims datasets that provide information regarding patients' clinical characteristics. We hypothesized that there exists similarity in the characteristics (i.e., patients' clinical and socio-demographic factors) associated with a group of patients dependent on opioids. Identifying those patients and their associated characteristics allows for the development of a prediction model that can support physicians to determine the likelihood of patients' future dependence on opioids.
ii. We demonstrate that apart from logistic regression, other cutting-edge machine learning algorithms superiorly predict the likelihood of opioid dependency using a large analytic sample that is class imbalanced and entails skewness and sparsity of features. Along with higher predictive power we identify risk factors associated with opioid dependency that will add knowledge to the existing literature. In addition, our study opens an avenue for researchers and healthcare practitioners to develop predictive analytics for other healthcare applications leveraging advanced machine learning techniques and healthcare claims datasets.

The rest of the paper is organized as follows: Section 2 presents the previous research that used administrative data in an assortment of healthcare applications along with the prediction of opioid dependency and also summarizes the state-of-the-art data mining and machine learning algorithms used for predictive analytics. Section 3 outlines and describes the proposed predictive analytics framework. Section 4 demonstrates the results and provides a discussion of the findings. Finally, Section 5 concludes with a summary of the work conducted in this research along with some limitations and future research directions.

2. Review of related research

This section provides a summary of previous research that has used healthcare administrative datasets in a variety of healthcare applications. In addition, we also summarize the existing research that has utilized claim-based datasets to address the opioid crisis. Furthermore, a brief overview of state-of-the-art data mining and machine learning algorithms that were used for outcome prediction in several healthcare sectors is introduced.

2.1. Related work leveraging healthcare administrative claim data sets

Prior research [14] suggests that studies based on de-identified administrative claim datasets have become increasingly common since the early 2000s. Such claims-based datasets have proven to be valuable in an assortment of health services applications including but not limited to drug utilization pattern research [15-18], disease burden estimations for diseases such as hypertension, diabetes and different heart conditions [19-21], quality-of-care assessment [22, 23], and health policy evaluations [24, 25]. In the past 10 years, detection of adverse drug events [26, 27] and outcome prediction [28] have emerged as a promising and potential application of claims-based research, revealing information about prescription patterns, efficacy, and safety.

Researchers used de-identified administrative databases from Aetna, a health insurance company, to assess the effect of varying opioid prescribing patterns following surgical procedures on the dependence, overdose, or abuse in a group of naïve opioid patients. [18], They found that duration of opioid prescriptions rather than the dosage is more strongly associated with eventual misuse in the early postsurgical period. However, their study cohort was limited to postsurgical patients. Likewise, prior research [7, 29, 30] suggests that the risks associated with opioid overdose is driven by the dosage (greater than 120mg morphine equivalent dosing per day), duration of chronic opioid use, and type of opioid medications (e.g., short or long acting). Additionally, patients' demographics (e.g., age, ethnicity, geographic regions) [31], prior clinical history of mental illness [9, 32], non-opioid substance use disorder [9], tobacco abuse [32], and non-dependent alcohol abuse [9, 32, 33] were also considered as potential risk indicators for opioid overdose. Doctor shopping (receiving opioid prescriptions from more than two physicians) and pharmacy shopping (purchasing opioids from more than three pharmacies) were also proven as significant risk predictors in the existing literatures [31, 34-36].

The current state-of-the-art of relevant research based on claims datasets have focused on specific cohorts of opioid users and/or investigated a selected set of risk factors, which is arguably narrow in scope. This leads to the following major limitations of existing studies: (1) ***lack of generalizability*** in identifying opioid utilization patterns among a broader segment of naïve opioid users, and (2) ***lack of empirical***

knowledge to comprehensively assess how several known or suspected risk factors influence patents' progression towards opioid dependency.

2.2 Review of data mining and machine learning algorithms used in healthcare risk prediction

A great deal of prior research demonstrates the successful application of machine learning techniques for developing predictive models that leverage large-scale retrospective datasets across several healthcare domains. Such studies include but are not limited to predicting mortality of intensive care unit (ICU) patients [37-41], sepsis related organ failure assessments [42-44], breast cancer survivability [45], risk of 30-day hospital readmission for patients with heart failure, pneumonia, serious mental illness (SMI) [46, 47], generating treatment plan for diseases such as ulcers [48], and depressions [49], and predicting risk of opioid overdose [7-9, 36, 50, 51] etc.

Previous research [52, 53] for mortality prediction of ICU patients has found superior performance for artificial neural networks (ANN) over regression-based predictive modeling approaches, while another groups of studies [54-56] indicates that both of these two predictive algorithms perform equally for mortality prediction. Some other groups of studies [42, 45, 57, 58] have adopted decision tree and support vector machine algorithms and their findings showed that these two algorithms outperformed artificial neural networks and logistic regression techniques. Improved and better predictive accuracy was found for support vector machine compared to acute physiology and chronic health evaluation-II (APACHE II) in [42], yet C5.0, a variant of decision tree algorithm, exhibited the best performance followed by support vector machine, APACHE III and artificial neural network [57].

Overall, the evidence for the predictive performance of different state-of-the-art machine learning techniques reveal the following two issues: (1) no single classification technique invariably shows superior performance over all other techniques, and (2) performance varies based on the study populations, the selected features, and outcome measure of interest. For instance, the descriptive modeling characteristics of decision tree algorithm makes it a better choice to explain the hidden implications over artificial neural network, which oftentimes fails to explain the association between predictors and outcome of interest [59]. From a big-data viewpoint, decision tree, random forest, artificial neural network, and support vector machine learning showed the ability to handle large data samples and incorporate prior knowledge into predictive analysis [60].

2.3. Existing model for the prediction of opioid abuse, overdose or dependence

In the literature, there are several cases where opioid misuse was examined and risk factors were identified. For instance, in one study [11], logistic regression and split sample validation were used to examine the differential effect of risk factors of drug overdose between males and females. It was found that substance abuse was the greatest predictor of opioid overdose for both male and female. However, the higher odd ratio related to male indicated stronger association between opioid overdose and male patients. These results were reaffirmed by other studies [7, 9] using logistic regression techniques, which found that males were more likely to be addicted to opioids. Researchers [61] studied persistent opioid use after minor and major surgeries using a nationwide insurance claims dataset, which found that substance abuse was a major risk predictor. Another study [12] using a retrospective observational cohort based on medical and pharmacy claims from a nationally representative sample applied logistic regression and found that the initial opioid regimen is a strong predictor of chronic opioid therapy.

In predicting high risk of opioid overdose among non-cancer pain patients with chronic opioid therapy , logistic regression achieved 79% accuracy [62]. A cox regression model [63] was applied to a retrospective cohort to predict which patients were suitable for naloxone prescriptions, and was able to distinguish between high-risk and low-risk patients with a predictive accuracy of between 66-82%. Another study [50] used stepwise logistic regression on retrospective claims data to identify key risk factors and create a predictive model for opioid abuse.

After a comprehensive review of state-of-the-art research on the applicability of machine learning techniques to predict risk of opioid overdose, we identified the following limitations of previous research: (1) small sample sizes that did not support robust analytical files, (2) a narrow set of pre-selected clinical and socio-demographic factors for inclusion in the predictive model, and (3) only multi-variate logistic regression and cox regression based predictive analytics were used, which reported low predictive accuracy thus reducing the reliability and applicability of existing predictive models to diagnose opioid dependency. Although several cutting-edge machine learning techniques have been shown to achieve superior predictive performance while predicting risk scores in a variety of healthcare applications, research that leverages those advanced machine learning techniques to accurately predict the risk of opioid dependency is still in its infancy. These limitations of the existing studies create an avenue for further research to deploy state-of-the-art data mining and machine learning techniques on large-scale retrospective and administrative claims dataset and also to investigate their predictive power.

3. A framework for predicting opioid dependency using big-data analytics

In this section, we present our proposed framework to predict the likelihood of a patient's dependency on opioids. As previously mentioned, patients who are prescribed opioids for acute or chronic pain

management frequently become addicted. Along with the potency and duration of prescribed dosage there are also other factors entailing a patient's clinical history such as socio-economic and demographic characteristics, which together potentially lead a naïve opioid patient toward opioid dependence. Given variation in patient characteristics as mentioned above, there might be different paths that underlie the tendency or likelihood of a patient becoming dependent on opioids. We investigate such paths as constituting distinctive patient-level characteristics. Identifying the influential factors associated with patients who at some point abuse these drugs is critical to determine the risk of being dependent on opioids. We hypothesized that there exists similarity in these factors for patients who are dependent on opioids. Identifying and classifying such patients who have a similar profile based on clinical and socio-demographic factors would be an effective way to identify appropriate usage patterns for prescription opioids. This could potentially help clinicians understand the trajectory of how a naïve opioid user eventually develops opioid dependency. Thus, our proposed framework starts with determining the influential features associated with opioid dependency followed by the implementation and validation of cutting-edge machine learning algorithms as depicted in Figure 1 and Figure 2, respectively.

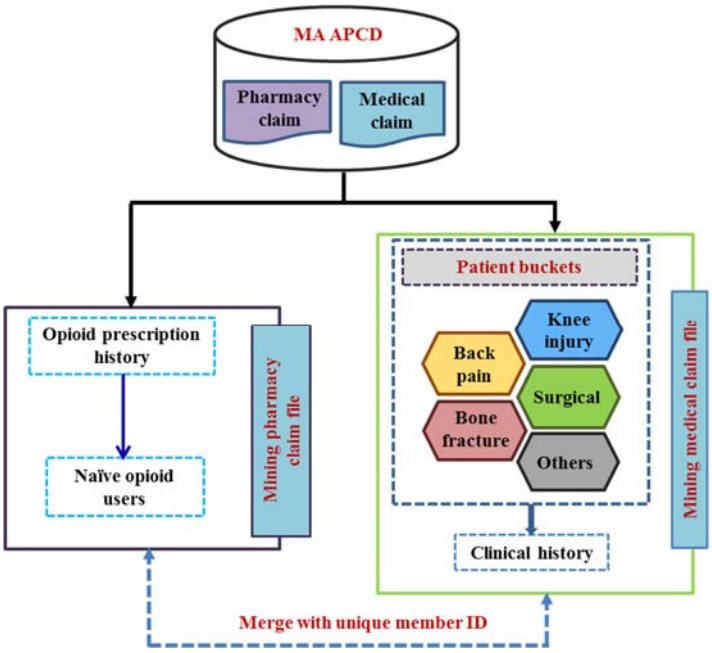

Figure 1. Outlining the development of the analytic file from the MA APCD

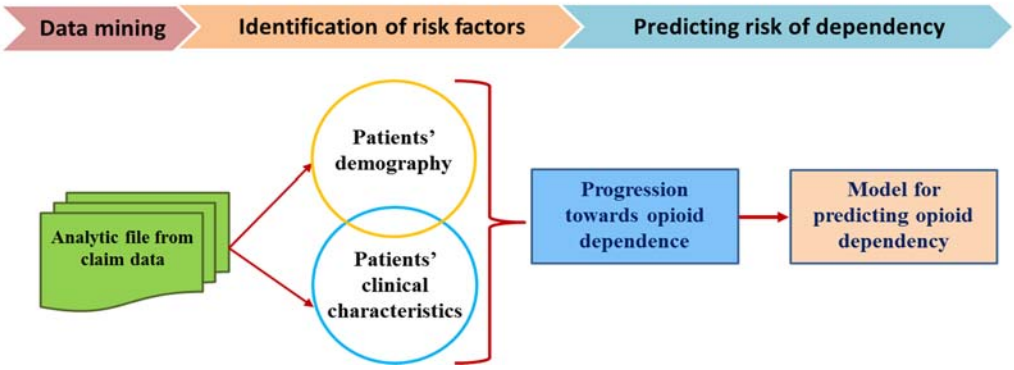

Figure 2. Outline of developing a predictive model for opioid dependency

3.1. Feature engineering to determine influential risk predictors associated with prescription opioid overdose.

Given that the success of an efficient risk prediction model depends on how the features are being presented, we transform the raw MA APCD into influential features that have significant association with opioid dependency. Engineering these factors is critical to help develop meaningful insight into how each unique naïve opioid patient is progressing towards opioid dependency and also to help improve the accuracy of the predictive algorithm with unseen/test data.

3.1.1. The Massachusetts All Payer Claim Datasets (MA APCD)

We use the *Massachusetts All Payer Claim Datasets (MA APCD)* [13] which we currently house at the Northeastern Center for Health Policy and Healthcare Research, and is overseen by the Center for Health Information and Analysis (CHIA). This unique database includes all medical claims, pharmacy claims, and member eligibility information associated with commercial insurance claims in Massachusetts between 2011 and 2015.

The dataset provides unique identifiers for all patients and types of providers (e.g., hospitals, physicians, nursing homes, rehabilitation facilities) that can be used to link claims for individuals across files. The pharmacy claims file contains data for approximately 470 million prescriptions and the medical claims file has approximately 1.63 billion claims for Massachusetts residents. The pharmacy claim files contains information on each prescription claim including the type of medication, the dosage, and the days supply. The medical claim files includes information pertaining to a patient's clinical condition (including principal diagnosis recorded as ICD-9 codes, services and procedures received, and payment (i.e., how much the provider received from the health plan for the services provided). Both the pharmacy and medical claims files can be linked to the member eligibility file that contains information on age, gender, insurance status (e.g., name and type of plan including co-pays and deductible), We hypothesize that understanding

the opioid usage pattern of naïve users and determining the impact of different factors on the outcome of interest (i.e., opioid dependency) are possible using the de-identified administrative claim datasets.

3.1.2. Dealing with missing values and data preprocessing

Age (categorical): Age for some opioid non-dependent patients were missing, which we replaced with mean age of all other non-dependent patients (after considering overall age distribution). We considered age as categorical feature instead of a continuous one. We divided age into five different buckets: i) 18-25, ii) 26-35, iii) 36-55, iv) 56-64, v) 65+ in an effort to understand whether or not a patient belonging to a certain age-bin is more likely to develop dependency than other.

Gender (categorical): Several records had gender values either missing or 'unknown'. The distribution of remaining data for male/female is ~50:50, so instead of imputing missing or unknown values as either male or female, we replaced them with 'unknown'. In other words, we have gender values as: male, female, or unknown.

ICD-codes (categorical): In the data, the primary diagnosis has been recorded as standard ICD-codes (which are alpha-numeric characters). As they are difficult to interpret in their original form, we replaced ICD-codes with their actual descriptions to make the data and results more interpretable.

3.1.3. Description of feature engineering procedure

We develop an initial ***analytic file*** comprised of potentially influential factors to be used for predicting risk of developing dependencies on opioids. These factors are ascertained before a patient develops or shows any sign of dependency or addiction on opioids. The underpinning idea was to extract potentially relevant socio-demographic characteristics and clinical history by using feature engineering techniques on the MA APCD.

Specifically, we identify the pattern of how a naïve opioid patient develops a dependency on opioids. Naïve opioid patients were identified from the pharmacy claims dataset. These are the patients who filed at least one claim that was identified as an opioid-related claim (identified using national drug codes, NDC), had no other opioid prescriptions, and were not diagnosed with an opioid dependency or one year prior to the index date. The index date was chosen as the most recent fill date of an opioid prescription during the study time frame.. For each naive opioid patient, we take the first opioid prescription and from the date of that prescription we track the patient from 6 to 12 months further in time in the medical claims dataset to check whether or not they had developed any sort of dependency on opioids—identified by ICD-9 diagnosis codes associated with drug dependence and overdose. If yes, then we consider the patient as opioid-dependent, and non-dependent otherwise. We use the term OD to refer to opioid-dependent patients

and ND to refer to non-dependent patients. For all such patients, we go one year back in time and gather that their medical history. Tracking of patients was censored when any incidence of opioid overdose was identified within the predefined time window.

3.1.3.1. Eliminating features with low variance

The final analytic file comprises approximately 600,000 patients and 12,000 features. Apart from age, sex, and zip code, all other features entail patients' clinical history extracted as ICD-9 codes. If a patient was diagnosed with a particular ICD-9 code, then we recorded that as n or 0 (where n is the number of times that patient was diagnosed for that particular ICD-9 code). This dimension made further analysis difficult, so we decided to eliminate less important features. Without training any model, we were required to eliminate features, so we chose the variance threshold (VT) method, which is a feature selection technique that removes features with low variance. This algorithm only looks at features (or predictors) and not the outcomes (or target). So, it is similar to an unsupervised machine learning technique. It removes features that have the same value for all the records in the sample.

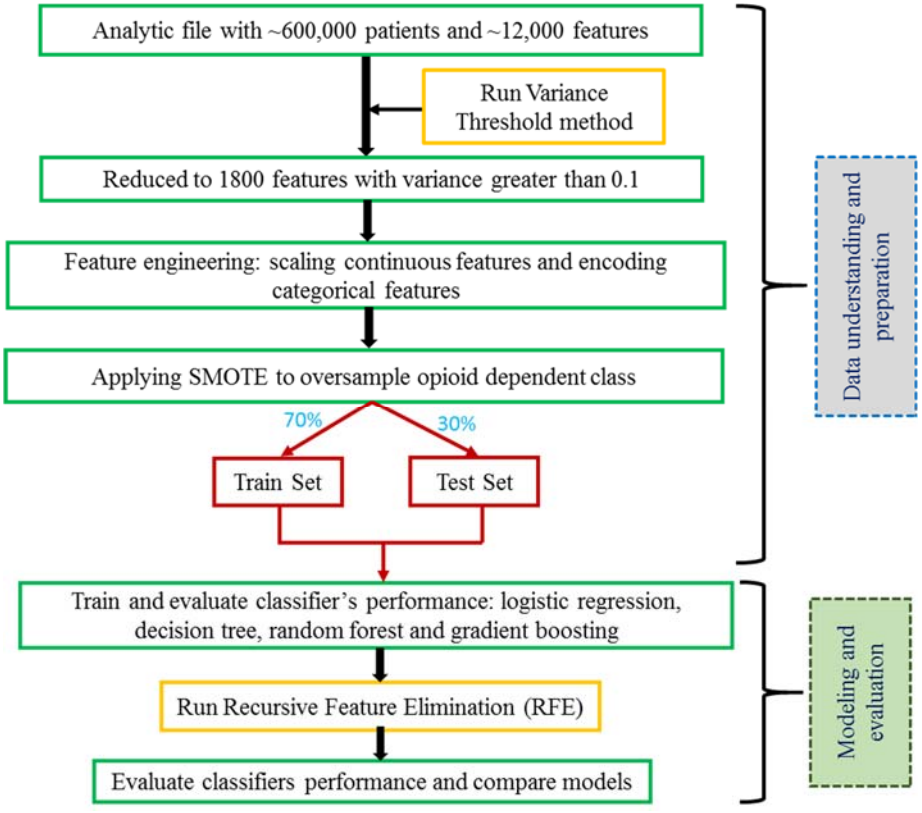

Figure 3. Data preparation, development and evaluation of predictive

By having multiple features in the data-set, we expect to capture certain variance in the data that can help us in predicting the outcome. We used the VT method, to calculate the variance of each column and

then removed the columns that had a variance less than the predefined threshold. As there does not exist a standard value that works for all types of datasets, we had to follow a trial-and-error approach to select an appropriate threshold. However, there exists some general guidelines as per scikit-learn's documentation [64]: Suppose that we have a dataset with Boolean features and we want to remove all features that are either one or zero (on or off) in more than 80% of the samples. Boolean features are Bernoulli random variables, and the variance of such variables is given by $Var(x) = p(1-p)$. In our data, we wanted to retain as many features as we could initially, so after certain trials we decided to remove features that have $\geq 90\%$ of values missing (or 0). Thus, our threshold becomes: 0.9*(1-0.9) = 0.09~0.1.

3.1.3.2. Retaining most important features

After VT, we had ~1800 features remaining, i.e., ~1800 features had variance ≥ 0.1. Our goal was to make the models more interpretable and computationally more efficient. To reduce features we used Recursive Feature Elimination (RFE) technique. As the name suggests, it removes (or eliminates) features recursively by pruning the original set of features from the model. RFE takes a pre-trained model (or estimator) that assigns weights to whole set of features (either as coefficients for linear model or feature importance for tree based model). In each iteration, the least important features are pruned based on weights. Byy doing this, it achieves the goal of getting a smaller set of features in each iteration and eventually end up with only the desired number of features. As such, we were required to decide how many features we want to prune at the end of each iteration and provide a stopping condition for the RFE to terminate in order to get a feasible result. Using cross-validation technique we decided to remove 10% of the features after each iteration and have the top 30 from 1800 features as a stopping criteria.

Another subtle consideration with RFE is that if we use it for different types of models such as logistic regression, decision tree, random forest etc., then it produces (or retains) different sets of features for different models when the algorithm terminates. For example, if we want to retain 50 features and we use RFE with logistic regression and random forest, we may end up with two different sets of 50 features from these two models. That is, these 50 features may or may not be same for two different models.

3.1.3.3. Handling class imbalance data

In the final analytic file the prevalence of OD patients was only ~1%. Selection of an appropriate performance metric was critical to evaluate the models' performance. To achieve better results, we implemented Synthetic Minority Oversampling Technique (SMOTE)—a technique to over-sample a minority class (OD patients). This technique generated 'synthetic' samples of a minority class (OD patients) and thus a similar distribution of ND and OD patients in our dataset.

3.2. Development and implementation of machine learning algorithms to facilitate the proposed risk prediction model

We investigate and evaluate the performance of state-of-the-art machine learning algorithms to precisely classify the individuals at risk for opioids. We framed this problem as a binary classification problem *(i.e., target-variable is: '0' if patient was not dependent and '1' if dependent on opioids*). As we did not have labeled data, i.e., we did not have a *target-variable* that would tell us if a person became opioid-dependent, we were required to infer the labels using clinical history associated with opioid dependence and (or) overdose from the medical claims file. It was accomplished while preparing the final analytic file with essential features associated with opioid dependency.

Looking into the final analytic file, we observed big sparse features (high-dimensional and non-linear feature space) that may cause logistic regression based traditional classifiers to fail in terms of learning features or overfitting issues. As previously discussed, no single algorithm is always better than others, we intend to adopt multiple predictive modeling approaches to present an analytic pipeline for predicting risk of opioid dependency. More specifically, we intend to assess the performance of traditional logistic regression technique and also tree based classifiers (decision tree, random forest, and gradient boosting).

At first, we split the analytic file into train-test set (70/30 split). As the dataset is highly imbalanced (i.e., prevalence of OD class is ~1%), we perform SMOTE to balance the training data and get a 50/50 distribution of ND and OD class. As a next step, we train different models as mentioned before. After training each model on an entire training data set, we evaluate performance on both train and test data by using: precision, recall, f-score, and area under receiving operating characteristic (AUC) curve. We then further eliminate features via RFE technique as discussed before, and end up with 30 features that contain the most amount of predictive power. Finally, we train the model again with remaining most important features and evaluate performance as explained earlier. These steps are outlined in Figure 3.

4. Results and discussions

4.1. Comparison of different models in predicting risk of opioid dependency

In this section, we present a comparative analysis of four predictive modeling approaches based on some performance measures. We also describe the influential features that are significantly important for predicting opioid dependency. For the tree-based models, we present these features in the Appendix. As mentioned in the previous section, the prevalence of OD patients is only ~1%. Therefore, accuracy alone cannot be considered as a reliable performance measure and choosing an appropriate evaluation metric is critical for selecting the best predictive model. As such, for each model, we determined precision, recall, F-1 score and AUC, which is presented in Table 1 and Table 2, and graphical comparison was shown in

Appendix (Figure 2 and Figure 3). The change of model recall value after VT and RFE is shown in Figure 4. We chose recall and AUC considering the following two reasons:

i. We intended to optimize the model's best 'recall', meaning that we want to identify as many OD patients as we can and are willing to accept misclassifications only when an 'ND patient is identified as OD'. Identifying an OD patient is more critical than misclassifying an ND patient, because if an OD patient is misclassified, then it is likely that this patient goes without proper medication management and becomes dependent on opioids, potentially leading to drug overdose. On the other hand, if we misclassify ND as OD then at most the medical practitioner may end up with prescribing some alternative treatment which is likely to be even less addictive.

ii. AUC is very effective in indicating model's effectiveness to identify a rare class from the prevalent one. So, we also use that metric to evaluate model's performance.

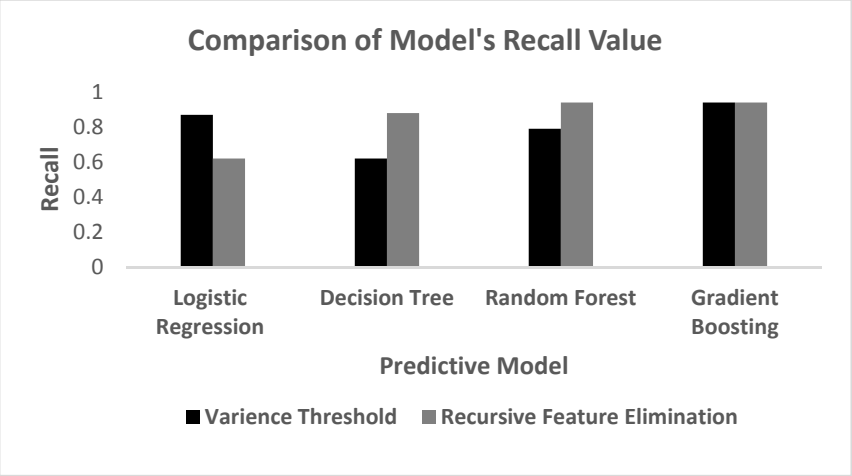

Figure 4. Comparison of recall value for different predictive models

Based on these two performance measures we present a comparative analysis across all four predictive models in the following sections. Figure 5 shows the area under the AUC curves for all four models. The value of this metric varies between 0 and 1, and the higher the values the better the model's performance. In other words, we expect the curve to reach top-left corner and thus have a value closer to 1. The AUC in case of test data are highest for random forest and gradient boosting for any value of false positive and true positive rates, and it is slightly lowest for decision tree model. Noticeably, we have observed that the AUC for tree based model significantly outperforms logistic regression model.

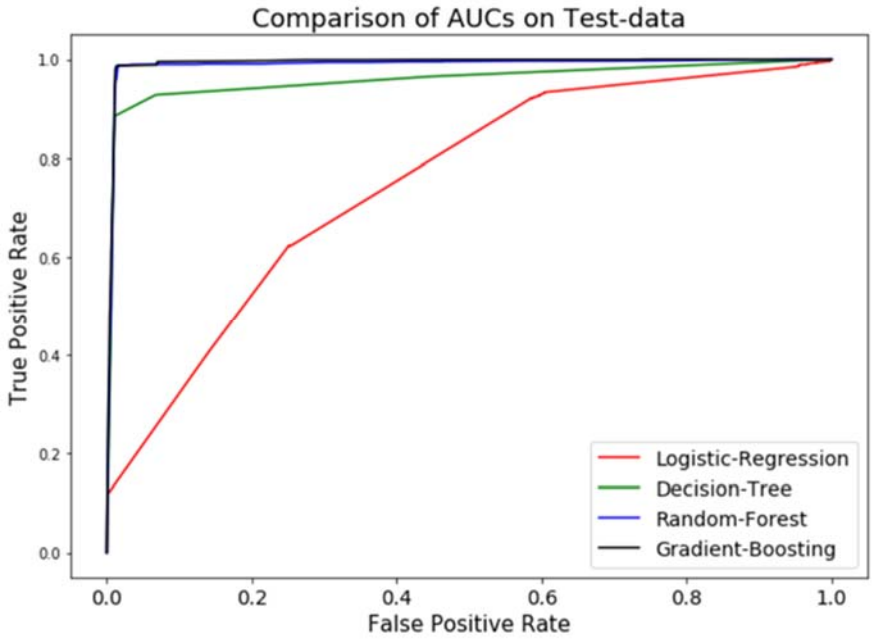

Figure 5. Comparison of ROC for different predictive models

Table 1
Performance of different models on training data

Model	Step	Class	Precision	Recall	F-1 score	AUC
Logistic regression	VT	ND	0.94	0.99	0.96	0.962

		OD	0.99	0.94	0.96	
	RFE	ND	0.75	0.75	0.75	0.746
		OD	0.75	0.75	0.75	
Decision tree	VT	ND	1	1	1	1
		OD	1	1	1	
	RFE	ND	0.98	0.99	0.99	0.986
		OD	0.99	0.98	0.99	
Random forest	VT	ND	0.98	0.99	0.98	0.985
		OD	0.99	0.98	0.98	
	RFE	ND	1	0.99	0.99	0.992
		OD	0.99	1	0.99	
Gradient boosting	VT	ND	1	0.99	0.99	0.993
		OD	0.99	1	0.99	
	RFE	ND	1	0.99	0.99	0.993
		OD	0.99	1	0.99	

Abbreviations: VT: Variance Threshold; RFE: Recursive Feature Elimination; ND: Non-dependent; OD: Opioid Dependent; AUC: Area Under Curve

Table 2
Performance of different models on test data

Model	Step	Class	Precision	Recall	F-1 score	AUC
Logistic regression	VT	ND	1	0.99	0.99	0.929
		OD	0.44	0.87	0.59	
	RFE	ND	0.99	0.75	0.85	0.685
		OD	0.03	0.62	0.06	
Decision tree	VT	ND	1	0.99	0.99	0.807
		OD	0.58	0.62	0.6	
	RFE	ND	1	0.99	0.99	0.937
		OD	0.51	0.88	0.65	
Random forest	VT	ND	1	0.99	0.99	0.889
		OD	0.41	0.79	0.54	
	RFE	ND	1	0.99	0.99	0.965
		OD	0.51	0.94	0.66	
Gradient boosting	VT	ND	1	0.99	0.99	0.965
		OD	0.52	0.94	0.67	
	RFE	ND	1	0.99	0.99	0.965
		OD	0.52	0.94	0.67	

Abbreviations: VT: Variance Threshold; RFE: Recursive Feature Elimination; ND: Non-dependent; OD: Opioid Dependent; AUC: Area Under Curve

4.1.1. Logistic regression

The logistic regression model performed well on training data with an AUC of 0.962, but its performance was not equally good on test-data. AUC dropped to 0.929, and the model's recall also dropped significantly (from 0.94 to 0.87). More importantly, when we eliminated features using RFE, performance

of the logistic regression model degraded significantly. AUC for training and test-data was 0.746 and 0.685, respectively. Similarly, a smaller value of model recall was observed both in case of training and test set (0.75 for training case and 0.62 for test case). Yet, we still get some meaningful insights from the model by understanding Odds Ratio (OR) of features.

The top-8 features that have the highest OR are: age 26 to 35 (5.14), poisoning by heroin (4.51), age 18 to 25 (3.33), age 36 to 55 (2.75), gender (2.44), poisoning by other medical substances and unspecified drugs (2.44) or alkaloids (2.04) or opiates and narcotics (2.09), and unspecified opioid abuse (1.58). Here, age 26 to 35 is a binary feature which has the highest odds ratio of 5.14. It implies that there exists a positive relationship between 'being in age bin 26 to 35' and opioid dependence tendencies. The patients falling into this age bin are 5.14 times more likely to develop dependency than those who don't fall into this age-bin. Similarly, patients in age bin 18 to 25 are 3.33 times more likely to develop dependency compared to others. Gender—represented as binary feature (0 for Female and 1 for Male) has an OR of 2.44, signifying that compared to female, male patients are 2.44 times more likely to develop dependency on opioids.

The rest of the top 8 features are discrete. OR for poisoning by heroin is 4.51, which means that if a patient was diagnosed once for poisoning by heroin, then it is 4.51 times more likely that this patient will develop opioid dependency as compared to patient who did not have the clinical history associated with this clinical diagnosis. For any discrete features representing clinical diagnosis in patient's history, the OR increases with the increase in number of times that a patient was diagnosed with that clinical characteristics according to OR^n, where n is the number of times a patient was diagnosed with a particular clinical characteristic.

4.1.2. Random forest

The random forest model performs quite well on training data after eliminating features based on a variance threshold. Its AUC for training data is 0.985. However, its performance on testing data is not equally good as AUC is 0.889, indicating that it's overfitting the training data. To diminish the effect of overfitting, we further optimized the model by removing redundant features using RFE. With 30 features in the data, AUC is 0.992 and 0.965 for training and test case, respectively. We can see the significant improvement in AUC for test case after eliminating redundant features.

We also get a high recall value of 0.99 for training data and for testing data it is 0.79, which is the ratio of the number of records classified as OD to the total number of records of OD. As mentioned earlier, it overfits the training data. But after RFE, recall for training and testing data are 1 and 0.94. It means that the random forest model effectively identifies OD patients with the redundant features being eliminated from the data. However, the drawback is its low precision, meaning that out of all the patients that the model

identified as OD, many of them were actually ND. Given our application, this is acceptable as it is very important to correctly identify an OD from ND, the reason of which was explained in section 4.1.

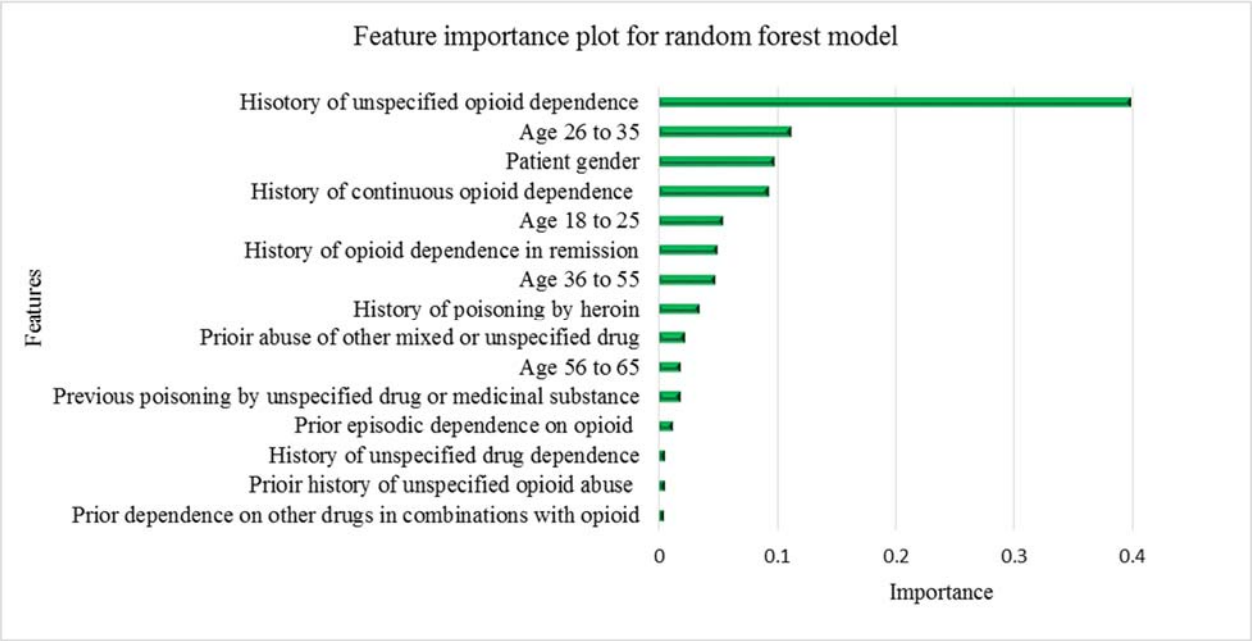

Figure 6. Feature importance plot for random forest model

After RFE, we are left with relatively more important features that improves model's performance entailing unspecified and continuous opioid dependence, age 26 to 35, gender, age 18 to 25, opioid type dependence in remission, age 36 to 55, poisoning by heroin, other unspecified drug abuse, age 55 to 65, long term use of other medications, and several other features. From the feature importance plot of reduced-random forest model, we see that along with previous diagnosis of opioid dependence, age 26 to 35 and male are very important predictors of opioid dependency. People in age group 26 to 35 are about twice as more likely to develop dependency than people in age group 18 to 25—a findings similar to logistic regression model as well.

4.1.3. Decision tree

We see similar trend in AUC and recall in the decision tree model as we saw in random forest. Initially it overfits the training data, but after RFE, it performs well on training as well as testing data. AUC on test-data after RFE is 0.937, which is slightly less than that of random forest model (0.965). Model's recall value on test data after RFE is 0.88, which is also less than that of random forest model (0.94).

Remaining important features in the decision tree model after RFE are mostly similar to those we observed from random forest model. However, this model gives more importance to following features as compared to random forest model: unspecified opioid type dependence, gender, and age 26 to 35.

4.1.4. Gradient boosting

The gradient boosting performs quite well on training as well as testing data after eliminating features using the VT technique. Its AUC for training data is 0.993 and for testing data it is 0.965. From the feature importance plot of gradient bosting, we see that it gives very high weightage to opioid type dependence unspecified. Other features which are considered very important by other models such as age, gender etc. are not important according to gradient bosting. Moreover, we also get a very high value of recall, 1 and 0.94 for training and test data, respectively. Precision and F-score for training data are 0.99, but the same for testing data are 0.52 and 0.67, respectively. From these, it's clear that it overfits the training data.

After RFE, with final 30 features remaining in the data, AUCs for training and testing data are same as before. We also see same values for precision, recall, and F-score. It means that our model effectively identifies OD patients. However, the drawback is its low Precision, implying that from all the patients that the model identified as OD, many of them were actually ND. Given our application, this is acceptable as it's very important to correctly identify an OD from ND.

In the feature importance plot after RFE, we see mostly the same features that we found before RFE. They include unspecified and continuous opioid dependence, age 26 to 35, gender, opioid type dependence in remission, poisoning by heroin, other unspecified drug abuse, and several other features with negligible weightage.

4.2. Discussion

This study examines the pattern of patients' long-term opioid usage in an effort to associate the risk factors that lead a naïve opioid user towards opioid dependency, and also leverage state-of-the-art machine leaning algorithms to predict the likelihood of such dependency. Unlike many existing studies on predicting opioid dependency, we allowed the selected predictive model to ascertain the risk factors associated with opioid dependency rather than building a model using pre-selected features. Often times, it is argued that administrative claims datasets lack adequate clinical details for developing efficient and reliable predictive models for patient However, our study demonstrates that such an approach is feasible and clinically significant in predicting potential opioid dependency by extracting influential clinical factors using large-scale healthcare administrative claim data.

Based on the comparison of several performance measures, our study showed that tree based models, in general outperform logistic regression model. Out of all three tree based models, random forest superiorly performs in classifying opioid dependent and independent patients. After RFE, we identified 30 features including patient's demographic and clinical characteristics from random forest model (Table 3 in

Appendix) that predict the likelihood of patients' dependency on opioids. The most important demographic features includes patients' age between 26 to 35 and 18 to 25, and male gender, a findings consistent with [50], which reported increased risk of opioid dependency for male patients. Among clinical factors our model determined that prior history of opioid dependence, poisoning by heroin, opioid abuse, and prior evidences of drug withdrawal adverse effect related to mental illness or psychotic disorder, abuse of non-opioid drugs or substances, poisoning by unspecified drugs or medical substances, and dependence on unspecified alcoholic substance are strong predictors of opioid dependency which supports several prior research [9, 65]. Other clinical features entail previous history of dependence on unspecified drugs and (or) combinations of opioid type drugs with other drugs, anxiety, depressive disorder, poisoning by opium (alkaloids), and long-term use of other medications add significantly valuable knowledge to what is already known regarding the risk factors of opioid dependence.

While several of our findings are noticeably consistent with prior clinical research examining the risk factors associated with opioid dependency [8, 9, 36, 50], the predictive power of our analysis substantially adds value to the existing literatures. Besides logistic regression technique, we also implemented other tree-based algorithms. Because our analytic sample entails a larger number of features with sparsity and skewness of features, it is perhaps not surprising that we found that tree-based models outperformed logistic regression in terms of AUC and recall value. As logistic regression belongs to a family of Generalized Linear Models (GLM), it can capture the linear relationship between predictors and outcomes, effectively. However, it is likely to fail in capturing some complex relationships. Because tree-based models are just structured hierarchy of rules to make predictions (not necessarily linear), they are robust to some above-mentioned data issues, and thus they outperform the logistic regression model. As such, we suggest that researchers studying healthcare claims data to predict opioid dependency should consider evaluating and implementing more than one classifier. Particularly, we recommend using the random forest model as it is evident from our analysis that the random forest model is able to identify opioid dependent patients with more accuracy as compared with logistic regression model while also efficiently determined the risk factors that have significant association with opioid dependency.

Additionally, the accuracy reported by c-statistic (alternative measure of AUC) of the existing predictive model might also be overestimated given the fact that the dataset was highly imbalanced with a reasonably smaller proportion of total study sample of patients identified as opioid abusers or dependent [7, 9]. Such imbalance issue, if not overcome properly could lead to an overfitted predictive model and biased estimation of accuracy value, which could potentially question the applicability and reliability of predictive model in such a highly critical application area. Our study, with the help of SMOTE tackled this issue and achieved higher Recall and AUC value, which overcome the limitation of biased learning from

imbalanced dataset. Hence, our investigation provides meaningful insights on risk factors associated with opioid dependency along with a predictive decision support system with higher reliability that can be instrumental for healthcare practitioners to help them identify patients at risk of developing dependency on opioid.

5. Conclusions, limitations and future research directions

In this study, we provide a machine learning framework that is capable of leveraging large-scale healthcare administrative claims data to reveal the underlying factors associated with opioid dependency. The proposed framework intends to serve as a decision support system for healthcare practitioners or physicians at the point of care or initial opioid prescription to help identify the patients at risk of future opioid dependency. To the best of our knowledge, this is the first study that utilize the MA APCD to develop a predictive analytics framework, particularly focusing on prescription opioid users to minimize the risk of being dependent on opioids—a critical issue that has been drowning the entire U.S, potentially led to a situation of addiction or overdose epidemic. While prior research largely dependent on the logistic regression based predictive technique suffered from overestimated accuracy due to imbalance data, our study tackled this issue utilizing SMOTE, and proposed a set of tree based predictive models that demonstrated superior predictive performance over traditional logistic regression model. Unlike traditional approach, we started with patients' entire medical history and allowed the model to determine the set of influential features that are significant predictors of such opioid dependency. We utilize variance threshold and recursive feature elimination technique in an effort to enhance model computational efficiency and interpretability. Moreover, we are able to show that along with higher predictive performance, random forest model determined the risk factors, some of those were consistent with prior findings of clinical literature. Thus, our study enables researchers and healthcare decision makers to utilize machine learning approach on healthcare claims datasets, and provide with the ability to predict potential opioid dependency.

We note several limitations of our study. It was not possible to track patients' clinical history before 2011 due to the unavailability of the data. Because data were not available beyond 2015, we were not able to investigate how many of them were diagnosed with opioid overdose in the future. As such, we had to exclude years 2011 and 2015 from being considered as an index year. In addition, we could not identify the opioid dependent patients in year 2014, as pursuant to federal policy, CHIA was required to remove all medical claims for drug dependence after 2013. Another limitation has to do with the selection of four (although well-known) out of the many other predictive algorithms.

Future study will extend implementing the current framework on claims dataset coming from other states that are highly afflicted with opioid overdose epidemic. Such analysis will also take into account

other predictive algorithms to investigate and compare their predictive power with current findings. We also aim to develop a user interface so that physicians can utilize this as ready-to-use tool, which will generate a risk score for an individual before prescribing opioids once the required information associated with risk factors are provided.

Appendix

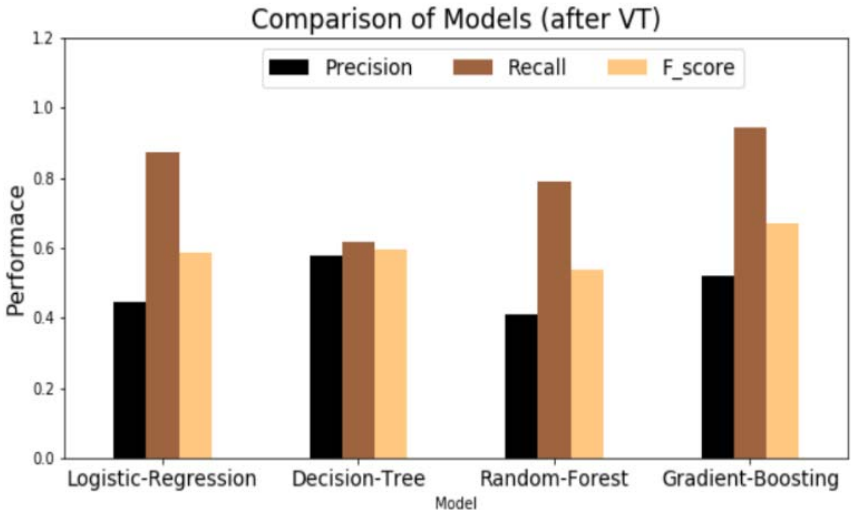

Figure 8. Comparison of Models after Variance Threshold method

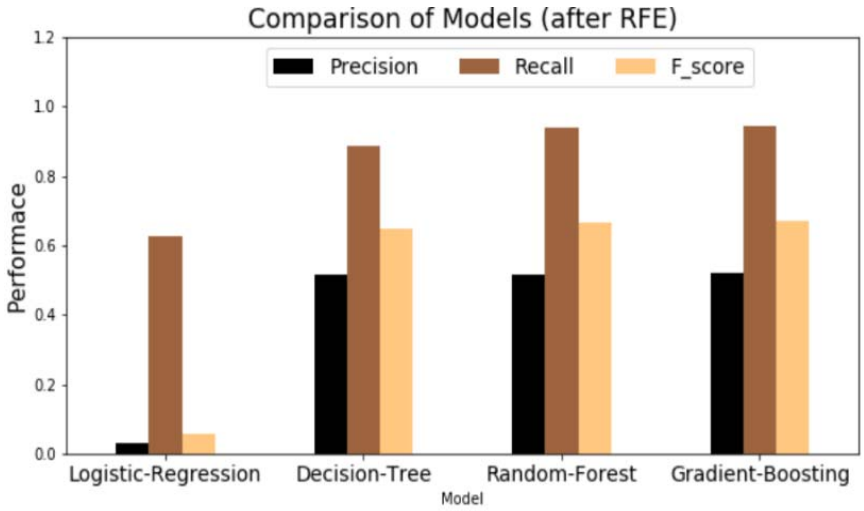

Figure 8. Comparison of Models after Recursive Feature Elimination

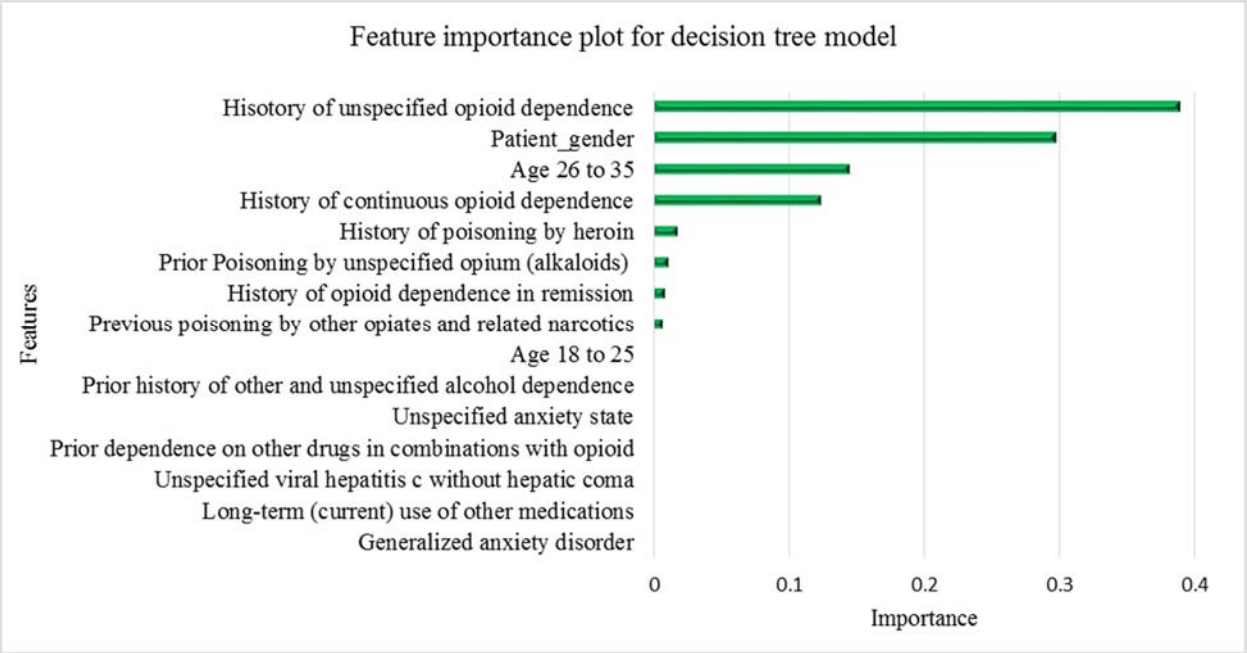

Figure 7. Feature importance plot for decision tree model

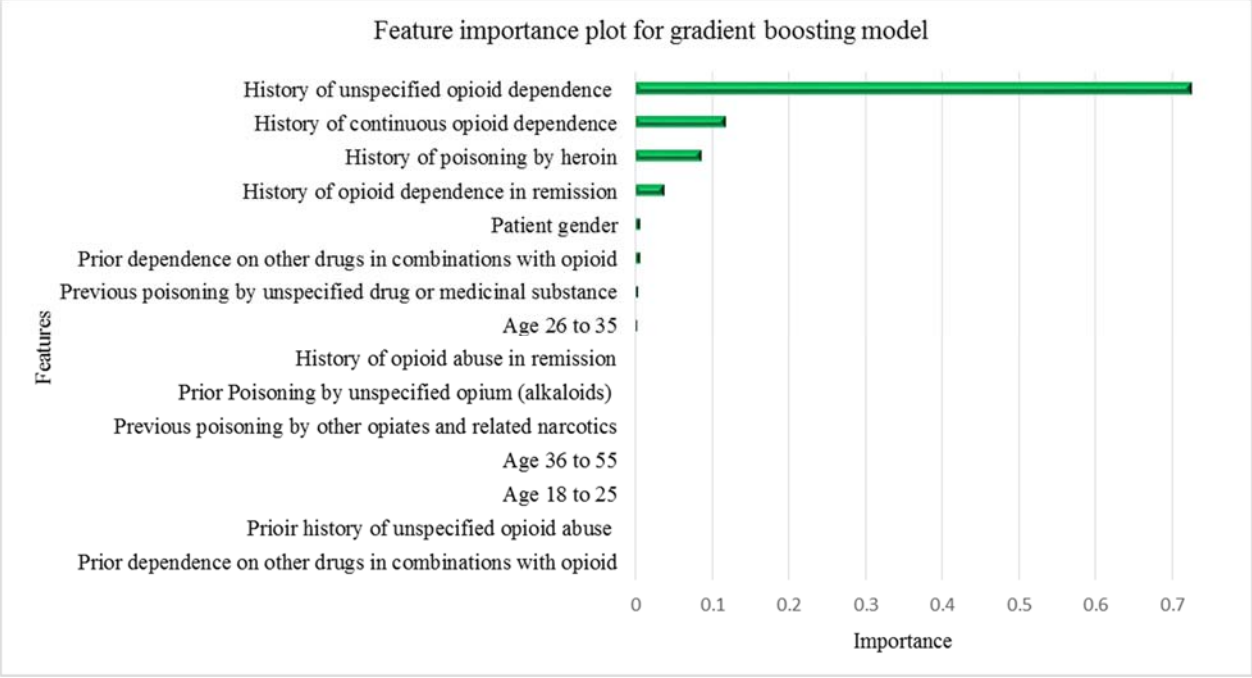

Figure 8. Feature importance plot for gradient boosting model

Table 3
Importance features for Random Forest model

Age 18 to 25
Patient gender
Need for prophylactic vaccination and inoculation against influenza
Age 26 to 35
Age 36 to 55
Drug withdrawal
Unspecified episodic mood disorder
Previous poisoning by unspecified drug or medicinal substance
Anxiety state unspecified
Age 56 to 65
Age 65+
History of unspecified alcohol dependence
Other screening mammogram
Combinations of drug dependence excluding opioids
History unspecified of opioid abuse
Routine general medical examination at a health care facility
Depressive disorder not elsewhere classified
Poisoning by opium (alkaloids) unspecified
History of unspecified drug dependence
Prior dependence (unspecified) on other drugs in combinations with opioid
Laboratory examination (unspecified)
Prior abuse of other mixed or unspecified drug
Prior dependence (continuous) on other drugs in combinations with opioid
History of opioid dependence in remission
History of episodic opioid dependence
Long-term (current) use of other medications
History of continuous opioid dependence
History of unspecified opioid type dependence
History of poisoning by heroin

Table 4
Important features from Decision Tree model

Generalized anxiety disorder
Long-term (current) use of other medications
Unspecified viral hepatitis c without hepatic coma

Prior dependence (unspecified) on other drugs in combinations with opioid
Anxiety state unspecified
Prior history of other and unspecified alcohol dependence
Patient gender
Age 18 to 25
Age 26 to 35
Unspecified episodic mood disorder
Prior poisoning by opium (alkaloids) unspecified
Prior dependence (continuous) on other drugs in combinations with opioid
Other malaise and fatigue
Routine general medical examination at a health care facility
History of poisoning by unspecified drug or medicinal substance
History of unspecified drug dependence
History of poisoning by other opiates and related narcotics
History of opioid dependence in remission
Fever unspecified
Routine gynecological examination
History of episodic opioid dependence
Chest pain unspecified
History of unspecified opioid abuse
History of continuous opioid dependence
History of opioid abuse in remission
History of Other mixed or unspecified drug abuse
History of unspecified opioid dependence
History of poisoning by heroin
Cough
Patient zip codes

Table 5
Important features from Gradient Boosting model

Age 36 to 55
Depressive disorder not elsewhere classified
Unspecified otitis media
Other disease of nasal cavity and sinuses
Esophageal reflux
Chronic rhinitis
Patient gender

Age 18 to 25
Other acute reactions to stress
Age 26 to 35
Age 56 to 65
Unspecified disorder of thyroid
History of opioid abuse in remission
History of opioid abuse in remission
History of unspecified opioid abuse
Prior dependence (continuous) on other drugs in combinations with opioid
Prior dependence (unspecified) on other drugs in combinations with opioid
History of episodic opioid dependence
History of continuous opioid dependence
History of unspecified opioid dependence
Prior Poisoning by unspecified opium (alkaloids)
History of unspecified mood disorder
History of poisoning by heroin
Previous poisoning by other opiates and related narcotics
History of poisoning by unspecified drug or medicinal substance
Long-term (current) use of other medications
Pain in joint lower leg
Lack of coordination
Patient Zip codes

References

[1] CDC. (2017, September 15). *Opioid Overdose, Understanding the Epidemic* Available: https://www.cdc.gov/drugoverdose/epidemic/index.html

[2] CDC. (2017, September 15). *Opioid Overdose, Drug OVerdose Death Data*. Available: https://www.cdc.gov/drugoverdose/data/statedeaths.html

[3] C. R. Chapman, D. L. Lipschitz, M. S. Angst, R. Chou, R. C. Denisco, G. W. Donaldson*, et al.*, "Opioid Pharmacotherapy for Chronic Non-Cancer Pain in the United States: A Research Guideline for Developing an Evidence-Base," *The Journal of Pain,* vol. 11, pp. 807-829, 2010/09/01/ 2010.

[4] CDC. (October 7). *CDC Guidelines for Prescribing Opioids for Chronic Pain*. Available: https://www.cdc.gov/drugoverdose/pdf/Guidelines_At-A-Glance-508.pdf

[5] (September 25, 2018). *THE MASSACHUSETTS OPIOID EPIDEMIC - A data visualization of findings from the Chapter 55 report*. Available: http://www.mass.gov/chapter55/

[6] T. Jones and T. Moore, "Preliminary data on a new opioid risk assessment measure: the Brief Risk Interview," *J Opioid Manag,* vol. 9, pp. 19-27, 2013.

[7] T. Ciesielski, R. Iyengar, A. Bothra, D. Tomala, G. Cislo, and B. F. Gage, "A tool to assess risk of de novo opioid abuse or dependence," *The American journal of medicine,* vol. 129, pp. 699-705. e4, 2016.

[8] T. J. Ives, P. R. Chelminski, C. A. Hammett-Stabler, R. M. Malone, J. S. Perhac, N. M. Potisek*, et al.*, "Predictors of opioid misuse in patients with chronic pain: a prospective cohort study," *BMC health services research,* vol. 6, p. 46, 2006.

[9] J. B. Rice, A. G. White, H. G. Birnbaum, M. Schiller, D. A. Brown, and C. L. Roland, "A model to identify patients at risk for prescription opioid abuse, dependence, and misuse," *Pain Medicine,* vol. 13, pp. 1162-1173, 2012.

[10] T. R. Hylan, M. Von Korff, K. Saunders, E. Masters, R. E. Palmer, D. Carrell*, et al.*, "Automated prediction of risk for problem opioid use in a primary care setting," *The Journal of Pain,* vol. 16, pp. 380-387, 2015.

[11] Y. Liang, M. W. Goros, and B. J. Turner, "Drug overdose: differing risk models for women and men among opioid users with non-cancer pain," *Pain medicine,* vol. 17, pp. 2268-2279, 2016.

[12] J. D. Thornton, N. Dwibedi, V. Scott, C. D. Ponte, D. Ziedonis, N. Sambamoorthi*, et al.*, "Predictors of Transitioning to Incident Chronic Opioid Therapy Among Working-Age Adults in the United States," *American health & drug benefits,* vol. 11, p. 12, 2018.

[13] (July 5). *Massachusetts All Payer Claim Datasets (MA APCD)*. Available: http://www.chiamass.gov/ma-apcd/

[14] F. Kari, B. Bryan, and J. Paul, "The use of claims data in healthcare research," *The Open Public Health Journal,* vol. 2, 2009.

[15] L. Kotzan, N. Carroll, and J. Kotzan, "Influence of age, sex, and race on prescription drug use among Georgia Medicaid recipients," *American Journal of Health-System Pharmacy,* vol. 46, pp. 287-290, 1989.

[16] A. Melander, K. Henricson, P. Stenberg, P. Löwenhielm, J. Malmvik, B. Sternebring*, et al.*, "Anxiolytic-hypnotic drugs: relationships between prescribing, abuse and suicide," *European journal of clinical pharmacology,* vol. 41, pp. 525-529, 1991.

[17] H. S. Glauber and J. B. Brown, "Use of health maintenance organization data bases to study pharmacy resource usage in diabetes mellitus," *Diabetes Care,* vol. 15, pp. 870-876, 1992.

[18] G. A. Brat, D. Agniel, A. Beam, B. Yorkgitis, M. Bicket, M. Homer*, et al.*, "Postsurgical prescriptions for opioid naive patients and association with overdose and misuse: retrospective cohort study," *bmj,* vol. 360, p. j5790, 2018.

[19] P. L. Hebert, L. S. Geiss, E. F. Tierney, M. M. Engelgau, B. P. Yawn, and A. M. McBean, "Identifying persons with diabetes using Medicare claims data," *American Journal of Medical Quality,* vol. 14, pp. 270-277, 1999.

[20] L. Quam, L. B. Ellis, P. Venus, J. Clouse, C. G. Taylor, and S. Leatherman, "Using claims data for epidemiologic research: the concordance of claims-based criteria with the medical record and patient survey for identifying a hypertensive population," *Medical care,* pp. 498-507, 1993.

[21] J. R. Robinson, T. K. Young, L. L. Roos, and D. E. Gelskey, "Estimating the burden of disease: comparing administrative data and self-reports," *Medical care,* pp. 932-947, 1997.

[22] K. N. Lohr, "Use of insurance claims data in measuring quality of care," *International journal of technology assessment in health care,* vol. 6, pp. 263-271, 1990.

[23] H. M. Krumholz, Y. Wang, J. A. Mattera, Y. Wang, L. F. Han, M. J. Ingber*, et al.*, "An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure," *Circulation,* vol. 113, pp. 1693-1701, 2006.

[24] S. Schneeweiss, C. Dormuth, P. Grootendorst, S. B. Soumerai, and M. Maclure, "Net health plan savings from reference pricing for angiotensin-converting enzyme inhibitors in elderly British Columbia residents," *Medical care,* pp. 653-660, 2004.

[25] R. Tamblyn, R. Laprise, J. A. Hanley, M. Abrahamowicz, S. Scott, N. Mayo*, et al.*, "Adverse events associated with prescription drug cost-sharing among poor and elderly persons," *Jama,* vol. 285, pp. 421-429, 2001.

[26] J. S. Brown, M. Kulldorff, K. A. Chan, R. L. Davis, D. Graham, P. T. Pettus*, et al.*, "Early detection of adverse drug events within population-based health networks: application of sequential testing methods," *Pharmacoepidemiology and drug safety,* vol. 16, pp. 1275-1284, 2007.

[27] J. A. Berlin, S. C. Glasser, and S. S. Ellenberg, "Adverse event detection in drug development: recommendations and obligations beyond phase 3," *American journal of public health,* vol. 98, pp. 1366-1371, 2008.

[28] L. Cai, J. Lubitz, K. M. Flegal, and E. R. Pamuk, "The predicted effects of chronic obesity in middle age on medicare costs and mortality," *Medical care,* pp. 510-517, 2010.

[29] E. Hoffman, J. C. Watson, J. St Sauver, N. P. Staff, and C. J. Klein, "Association of long-term opioid therapy with functional status, adverse outcomes, and mortality among patients with polyneuropathy," *JAMA Neurology,* vol. 74, pp. 773-779, 2017.

[30] M. Miller, C. W. Barber, S. Leatherman, and et al., "Prescription opioid duration of action and the risk of unintentional overdose among patients receiving opioid therapy," *JAMA Internal Medicine,* vol. 175, pp. 608-615, 2015.

[31] M. S. Cepeda, D. Fife, W. Chow, G. Mastrogiovanni, and S. C. Henderson, "Opioid shopping behavior: how often, how soon, which drugs, and what payment method," *The Journal of Clinical Pharmacology,* vol. 53, pp. 112-117, 2013.

[32] K. Skala, L. Reichl, W. Ilias, R. Likar, G. Grogl-Aringer, C. Wallner*, et al.*, "Can we predict addiction to opioid analgesics? A possible tool to estimate the risk of opioid addiction in patients with pain," *Pain physician,* vol. 16, pp. 593-601, 2013.

[33] S. E. McCabe, J. A. Cranford, and C. J. Boyd, "The relationship between past-year drinking behaviors and nonmedical use of prescription drugs: prevalence of co-occurrence in a national sample," *Drug and alcohol dependence,* vol. 84, pp. 281-288, 2006.

[34] L. A. Schloff, T. S. Rector, R. Seifeldin, and J. D. Haddox, "Identifying controlled substance patterns of utilization requiring evaluation using administrative claims data," *American Journal of Managed Care,* vol. 10, pp. 783-790, 2004.

[35] A. J. Hall, J. E. Logan, R. L. Toblin, J. A. Kaplan, J. C. Kraner, D. Bixler*, et al.*, "Patterns of abuse among unintentional pharmaceutical overdose fatalities," *Jama,* vol. 300, pp. 2613-2620, 2008.

[36] A. G. White, H. G. Birnbaum, M. Schiller, J. Tang, and N. P. Katz, "Analytic models to identify patients at risk for prescription opioid abuse," *The American journal of managed care,* vol. 15, pp. 897-906, 2009.

[37] J. G. Le, P. Loirat, A. Alperovitch, P. Glaser, C. Granthil, D. Mathieu*, et al.*, "A simplified acute physiology score for ICU patients," *Critical care medicine,* vol. 12, pp. 975-977, 1984.

[38] D. Poole, C. Rossi, N. Latronico, G. Rossi, S. Finazzi, G. Bertolini*, et al.*, "Comparison between SAPS II and SAPS 3 in predicting hospital mortality in a cohort of 103 Italian ICUs. Is new always better?," *Intensive Care Medicine,* vol. 38, pp. 1280-1288, August 01 2012.

[39] A. L. Rosenberg, "Recent innovations in intensive care unit risk-prediction models," *Current opinion in critical care,* vol. 8, pp. 321-330, 2002.

[40] Y. Luo, Y. Xin, R. Joshi, L. A. Celi, and P. Szolovits, "Predicting ICU Mortality Risk by Grouping Temporal Trends from a Multivariate Panel of Physiologic Measurements," in *AAAI*, 2016, pp. 42-50.

[41] R. Pirracchio, M. L. Petersen, M. Carone, M. R. Rigon, S. Chevret, and M. J. van der Laan, "Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study," *The Lancet Respiratory Medicine,* vol. 3, pp. 42-52, 2015/01/01/ 2015.

[42] V. J. Ribas, J. C. López, A. Ruiz-Sanmartín, J. C. Ruiz-Rodríguez, J. Rello, A. Wojdel*, et al.*, "Severe sepsis mortality prediction with relevance vector machines," in *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2011, pp. 100-103.

[43] S. Q. Simpson, "New Sepsis Criteria: A Change We Should Not Make," *Chest,* vol. 149, pp. 1117-1118, 2016/05/01/ 2016.

[44] E. Gultepe, J. P. Green, H. Nguyen, J. Adams, T. Albertson, and I. Tagkopoulos, "From vital signs to clinical outcomes for patients with sepsis: a machine learning basis for a clinical decision support system," *Journal of the American Medical Informatics Association,* vol. 21, pp. 315-325, 2013.

[45] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine,* vol. 34, pp. 113-127, 2005/06/01/ 2005.

[46] D. Kansagara, H. Englander, A. Salanitro, D. Kagen, C. Theobald, M. Freeman*, et al.*, "Risk prediction models for hospital readmission: a systematic review," *Jama,* vol. 306, pp. 1688-1698, 2011.

[47] S.-C. Chin, R. Liu, and S. B. Roy, "PREDICTIVE ANALYTICS IN 30-DAY HOSPITAL READMISSIONS FOR HEART FAILURE PATIENTS," *Healthcare Analytics: From Data to Knowledge to Healthcare Improvement,* p. 439, 2016.

[48] I. Cho, I. Park, E. Kim, E. Lee, and D. W. Bates, "Using EHR data to predict hospital-acquired pressure ulcers: a prospective study of a Bayesian Network model," *International journal of medical informatics,* vol. 82, pp. 1059-1067, 2013.

[49] M. C. Klein and G. Modena, "Estimating mental states of a depressed person with Bayesian networks," in *Contemporary Challenges and Solutions in Applied Artificial Intelligence*, ed: Springer, 2013, pp. 163-168.

[50] R. Dufour, J. Mardekian, M. K. Pasquale, D. Schaaf, G. A. Andrews, and N. C. Patel, "Understanding predictors of opioid abuse: predictive model development and validation," *Am J Pharm Benefits,* vol. 6, pp. 208-216, 2014.

[51] D. C. Turk, K. S. Swanson, and R. J. Gatchel, "Predicting opioid misuse by chronic pain patients: a systematic review and literature synthesis," *The Clinical journal of pain,* vol. 24, pp. 497-508, 2008.

[52] R. Dybowski, V. Gant, P. Weller, and R. Chang, "Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm," *The Lancet,* vol. 347, pp. 1146-1150, 1996/04/27/ 1996.

[53] A. Nimgaonkar, D. R. Karnad, S. Sudarshan, L. Ohno-Machado, and I. Kohane, "Prediction of mortality in an Indian intensive care unit," *Intensive Care Medicine,* vol. 30, pp. 248-253, 2004/02/01 2004.

[54] G. Clermont, D. C. Angus, S. M. DiRusso, M. Griffin, and W. T. Linde-Zwirble, "Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models," *Critical care medicine,* vol. 29, pp. 291-296, 2001.

[55] L. Wong and J. Young, "A comparison of ICU mortality prediction using the APACHE II scoring system and artificial neural network," 1998.

[56] G. Doig, K. Inman, W. Sibbald, C. Martin, and J. Robertson, "Modeling mortality in the intensive care unit: comparing the performance of a back-propagation, associative-learning neural network with multivariate logistic regression," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 1993, p. 361.

[57] S. Kim, W. Kim, and R. W. Park, "A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques," *Healthc Inform Res,* vol. 17, pp. 232-243, 12/ 2011.

[58] L. Citi and R. Barbieri, "PhysioNet 2012 Challenge: Predicting mortality of ICU patients using a cascaded SVM-GLM paradigm," in *Computing in Cardiology (CinC), 2012*, 2012, pp. 257-260.
[59] A. Awad, M. Bader-El-Den, J. McNicholas, and J. Briggs, "Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach," *International journal of medical informatics,* vol. 108, pp. 185-195, 2017.
[60] G. Meyfroidt, F. Güiza, J. Ramon, and M. Bruynooghe, "Machine learning techniques to examine large patient databases," *Best Practice & Research Clinical Anaesthesiology,* vol. 23, pp. 127-143, 2009/03/01/ 2009.
[61] W. J. Brummett CM, Goesling J, Moser S, Lin P, Englesbe MJ, Bohnert ASB, Kheterpal S, Nallamothu BK., "New Persistent Opioid Use After Minor and Major Surgical Procedures in US Adults.," *JAMA surgery,* 2017.
[62] J. M. Glanz, K. J. Narwaney, S. R. Mueller, E. M. Gardner, S. L. Calcaterra, S. Xu*, et al.*, "Prediction Model for Two-Year Risk of Opioid Overdose Among Patients Prescribed Chronic Opioid Therapy," *Journal of general internal medicine,* pp. 1-8, 2018.
[63] T. E. King Jr and M. B. Wheeler, *Medical management of vulnerable and underserved patients: principles, practice, and populations*: McGraw-Hill Medical Publishing Division, 2007.
[64] s. learn. (October 15). *sklearn.feature_selection.VarianceThreshold*. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.VarianceThreshold.html
[65] M. J. Edlund, B. C. Martin, M.-Y. Fan, A. Devries, J. B. Braden, and M. D. Sullivan, "Risks for opioid abuse and dependence among recipients of chronic opioid therapy: results from the TROUP study," *Drug and alcohol dependence,* vol. 112, pp. 90-98, 2010.