Journal of Biomedical Informatics 116 (2021) 103725

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Predicting opioid overdose risk of patients with opioid prescriptions using electronic health records based on temporal deep learning

Xinyu Dong [a], Jianyuan Deng [b], Wei Hou [c], Sina Rashidian [a], Richard N. Rosenthal [d], Mary Saltz [b], Joel H. Saltz [b], Fusheng Wang [a,b,*]

[a] Department of Computer Science, Stony Brook University, Stony Brook, NY, United States
[b] Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY, United States
[c] Department of Family, Population and Preventive Medicine, Stony Brook University, Stony Brook, NY, United States
[d] Department of Psychiatry, Renaissance Stony Brook Medicine, Stony Brook, NY, United States

ARTICLE INFO

Keywords:
Opioid overdose
Opioid poisoning
Deep learning
Clinical decision support
Electronic health records
Long short-term memory

ABSTRACT

The US is experiencing an opioid epidemic, and opioid overdose is causing more than 100 deaths per day. Early identification of patients at high risk of Opioid Overdose (OD) can help to make targeted preventative interventions. We aim to build a deep learning model that can predict the patients at high risk for opioid overdose and identify most relevant features. The study included the information of 5,231,614 patients from the Health Facts database with at least one opioid prescription between January 1, 2008 and December 31, 2017. Potential predictors (n = 1185) were extracted to build a feature matrix for prediction. Long Short-Term Memory (LSTM) based models were built to predict overdose risk in the next hospital visit. Prediction performance was compared with other machine learning methods assessed using machine learning metrics. Our sequential deep learning models built upon LSTM outperformed the other methods on opioid overdose prediction. LSTM with attention mechanism achieved the highest F-1 score (F-1 score: 0.7815, AUCROC: 0.8449). The model is also able to reveal top ranked predictive features by permutation important method, including medications and vital signs. This study demonstrates that a temporal deep learning based predictive model can achieve promising results on identifying risk of opioid overdose of patients using the history of electronic health records. It provides an alternative informatics-based approach to improving clinical decision support for possible early detection and intervention to reduce opioid overdose.

1. Introduction

In 2017, opioid overdose (OD) resulted in 47,600 deaths in the United States [1], approximately 130 deaths per day. The total economic burden of prescription opioid overdose and use disorder is estimated to be $78.5 billion in the United States [2]. In response, health systems and researchers have developed models to identify and intervene with patients at high risk of opioid overdose. In particular, the availability of the history of electronic health records (EHR) of patients and large scale EHR databases provides a unique opportunity to build effective machine learning driven predictive models which are capable of identifying patients at high OD risk and revealing critical features for the prediction.

In recent years, deep learning models have become increasingly favored for tasks involving electronic health records, due to their capacity to process large-scale data with minimal feature engineering [3–6]. A deep neural network (DNN) model has proven effective for disease prediction and improving coding accuracy from multiple facilities [5,7]. Among deep learning models, sequential based networks, such as Recurrent Neural Network (RNN), are increasingly favored for modeling disease progression as they can better capture complex patterns in temporal dimensions [8]. Long Short Term Memory networks (LSTM) is a special type of RNN, capable of learning long-term dependencies. RNN models have generally shown superior performance for processing text data, for example, an RNN model using clinical notes was used for predicting chronic diseases [9], and bidirectional RNN and LSTM models were used to label clinical texts [10] through predictions. RNNs are also useful for predictive modeling with EHR data, for example, Doctor AI employed an RNN model to predict diagnoses and

* Corresponding author at: Department of Biomedical Informatics, Department of Computer Science, Stony Brook University, 2313D Computer Science, Stony Brook, NY 11794-8330, United States.
E-mail address: fusheng.wang@stonybrook.edu (F. Wang).

https://doi.org/10.1016/j.jbi.2021.103725
Received 29 January 2021; Accepted 22 February 2021
Available online 9 March 2021

Table 1
Demographics for OD and non-OD patients.

	Non-OD patients	OD patients	p-Value
Total Number	7,284,389	60,646	
Sex			
Male	3,091,295 (42.44%)	25,671 (42.33%)	0.7914
Female	4,191,440 (57.54%)	34,620 (57.09%)	(χ^2 test)
Other/Missing	1,654 (0.02%)	355 (0.02%)	
Age (First medical exposure to opioid medication)			
0–15	695,162 (9.5%)	1,708(2.8%)	<0.0001
16–25	927,086 (12.73%)	6,945(11.45%)	(χ^2 test)
26–35	1,149,275 (15.78%)	8,931(14.73%)	
36–45	974,900 (13.38%)	8,264(13.63%)	
46–55	1,113,903 (15.29%)	10,708(17.66%)	
56–65	1,021,676 (14.03%)	9,926 (16.37%)	
>66	1,310,694 (17.99%)	12,927 (21.31%)	
Missing	91,693 (1.3%)	1,237(2%)	
Race			
Caucasian	5,187,030 (71.2%)	47,712 (78.67%)	<0.0001
African	1,072,442 (14.72%)	7,165 (11.81%)	(χ^2 test)
Asian	117,912 (1.6%)	400 (0.65%)	
Hispanic	128,536 (1.76%)	624 (1.0%)	
Other	685,514 (9.4%)	4,559 (7.5%)	
Missing	92,955 (1.27%)	186 (0.3%)	
Marital Status			
Single	2,925,696 (40.16%)	23,627 (38.96%)	<0.0001
Married	2,962,332 (40.67%)	19,604 (32.33%)	(χ^2 test)
Divorced	485,553 (6.7%)	7,195 (11.86%)	
Widowed	453,605 (6.2%)	5,923 (9.77%)	
Other	457,203 (6.3%)	4,297 (7.1%)	
Encounters			
Avg Number (Average ± Standard Deviation)	15.50 ± 25.60	19.08 ± 33.64	<0.0001 (T-test)
No. of patients with less than 5 encounters	2,821,844 (38.74%)	22,341 (36.83%)	<0.0001 (χ^2 test)
No. of Encounters with labs	4.39 (28.3%)	5.4 (28.3%)	<0.0001 (χ^2 test)
No. of Encounters with medications	2.77(17.85%)	3.11 (16.3%)	<0.0001 (χ^2 test)
Lab Tests			
Average Number (Average ± Standard Deviation)	184.48 ± 510.54	307.57 ± 979.29	<0.0001 (T-test)
Missing Values	29.45(15.96%)	44.35(14.4%)	<0.0001 (χ^2 test)

medications based on EHR data [11].

In this work, we introduce a method for predicting opioid overdose risk among patients prescribed with opioids, using deep learning models trained from the patients' EHR data. In order to represent long-term temporal effects, we employed a sequential predictive model based on LSTM. Meanwhile, to interpret critical features for the prediction, we employed the permutation importance method [12] to rank the top features in the model. Our model took advantage of the rich information in the EHR history, including both clinical and demographic information.

2. Methods

In this retrospective cohort study, we used de-identified EHR data from Cerner Health Facts from January 1, 2008 to December 31, 2017. We identified all patients with opioid prescriptions and split them into two groups, one containing patients with a diagnosis of opioid poisoning, and the other containing patients without any diagnosis of opioid poisoning.

2.1. Study populations and data sources

In this retrospective cohort study, we used data from Cerner's Health Facts is one of the largest EHRs databases in the United States. Health Facts includes de-identified patient data from over 600 participating healthcare facilities. This database is structured based on patients' encounters, which contain diagnoses, procedures, patients' demographics, medication dosage and administration information, vital signs, laboratory test results, surgical case information, health systems attributes [13], and other clinical observations.

As patients with an opioid prescription were the focus of this study, we extracted all patients who received a prescription for opioid or opioid-related medications. We searched DrugBank 5.1.4 and included medications with the Anatomical Therapeutic Chemical (ATC) level 3 code 'N02A' or with 'opioid' as the drug's category description [14]. The resulting list included butorphanol, diamorphine, eluxadoline, oxycodone, oxymorphone, naloxone, tramadol, levacetylmethadol, pentazocine, hydromorphone, levorphanol, remifentanil, normethadone, opium, sufentanil, piritramide, tapentadol, morphine, codeine, dezocine, fentanyl, nalbuphine, meperidine, naltrexone, buprenorphine, methadone, hydrocodone, alfentanil, dihydrocodeine, diphenoxylate.

We followed procedures from Moore [15] to identify OD patients by opioid poisoning ICD codes (Supplementary Table 1). Since opioids are known to be effective for cancer pain [16], patients with cancer are likely to receive high doses of opioids. Since these patients might be misclassified as having OD, we chose to exclude them. We identified cancer diagnoses based on cancer related ICD-9 [17] and ICD-10 codes [18] (Supplementary Table 2).

Furthermore, we filtered both OD and non-OD patients, excluding those who were first exposed to opioids before age 16 or after age 66. 72.21% of non-OD patients and 79.00% of OD patients remained in the study dataset. Table 1 shows the summary of information about selected patients. The process of patients selection is shown in Supplementary Fig. 1.

Table 2
Prediction performance of different methods.

Prediction Model	Precision	Recall	F-1	AUCROC	Area Under Precision Recall Curve
Random Forest	0.7695 ± 0.0056	0.7055 ± 0.0038	0.7361 ± 0.0057	0.8167 ± 0.0037	0.7044 ± 0.0029
Decision Tree	0.7277 ± 0.0053	0.7047 ± 0.0029	0.7160 ± 0.0029	0.7892 ± 0.0048	0.6837 ± 0.0027
Logistic Regression	0.7539 ± 0.0029	0.6050 ± 0.0026	0.6647 ± 0.0025	0.7147 ± 0.0035	0.7116 ± 0.0029
Dense Neural Network	0.8006 ± 0.0052	0.7329 ± 0.0046	0.7683 ± 0.0027	0.8214 ± 0.0028	0.7507 ± 0.0045
LSTM Network	0.7884 ± 0.0054	0.7616 ± 0.0027	0.7798 ± 0.0060	0.8318 ± 0.0051	0.7710 ± 0.0031
Attention	0.8128 ± 0.0019	0.7512 ± 0.0012	0.7815 ± 0.0022	0.8449 ± 0.0024	0.7856 ± 0.0021

A total of 5,231,614 patients with an opioid prescription between 2008 and 2017 were identified in Health Facts. 44,774 patients came with an opioid poisoning diagnosis ("positive"), while 5,186,840 patients did not have any recorded opioid poisoning event ("negative"). We split the group of negative patients into 10 equal parts (518,684 patients each). Each negative patient portion was combined with the positive patients for an evaluation, randomly allocating 80% to the training set and 20% to the test set. We repeated the evaluation for all 10 parts, and the performance metric of a model is the average value.

2.2. Preparation of features

We extracted 1185 features, including 414 diagnosis codes features, 394 laboratory test features, 3 demographic features, 227 clinical events features, and 147 medications features (Supplementary Table 1). Features present in fewer than 1% of positive patients' records were excluded to minimize overfitting with sparse features.

Diagnoses and procedures are recorded using ICD-9 and ICD-10 codes in Health Facts. To avoid an inflated number of features from combining two versions of ICD codes, we translated ICD-9 codes to ICD-10 codes [15], and to reduce granularity, we truncated ICD-10 codes to the first 3 digits. For instance, about the diagnosis of burn of unspecified degree of trunk, unspecified site, the code is 942.00 in ICD-9, and T21.00 in ICD 10. If there is no according conversion, we kept the original ICD-9 code. Medications, recorded by National Drug Code (NDC) in Health Facts, were converted into level-3 ATC codes [19] to represent medication categories, and the dose quantity of each medication was also calculated as an additional feature. In order to make fair comparisons between different opioids with varying potencies, we converted the doses to morphine milligram equivalents (MME) [20,21], which has been used in the CDC opioid-prescribing guideline [22]. Each laboratory test is recorded as a numeric value, in addition to a description of the value indicating whether it is "higher than", "lower than", or "within" the normal range. We recorded the ratio of values higher or lower than normal to the total number of laboratory tests each patient received respectively. In addition, Health Facts has a special class of clinical information designated "clinical events", which includes vital signs, certain procedures, and personal variables not otherwise classified, such as smoking history, height, weight, and heart rate. These were added to the feature list as well. Demographics such as age, gender and race/ethnicity are also included in the feature space.

2.3. Feature matrix construction

To prepare the data for training, we normalized features with varying representations. For diagnosis codes, we used a binary representation to record the existence or absence of each diagnosis code in each encounter. Ages were segmented into sequential 10-year age groups beginning with 15–24. Race/Ethnicity were encoded by one-hot encoding, a common scheme for coding categorical features, which transforms a feature with n distinct values to n binary variables indicating the presence (1) or absence (0) of each category [23].

We represented a clinical event into various representations, as the value can be either binary or numeric. For instance, the presence or absence of smoking history was encoded as binary, whereas height in centimeters was recorded as numeric. If there were multiple values of a clinical event in an encounter, e.g., heart rate, we included the median, maximum, and minimum values in the feature space. We applied median imputation for filling missing values. Previous research comparing various imputation methods, when applied to tasks involving noisy EHR data, has found no significant difference in performance [5,7].

For feature matrix construction, we first identified the prediction target encounter. For positive cases, the prediction encounter was the first encounter with a diagnosis of opioid poisoning. For negative cases, the last encounter was used as the target encounter. Then, we constructed the feature vectors using the last 5 encounters that were at least 14 days prior to the prediction encounter. If a patient had fewer than 5 encounters for the prediction, we repeated the last available encounter to fill the gap. For instance, if a patient had only 3 encounters prior to the prediction encounter, then the feature matrix for the patient would be composed of one feature vector for the first encounter, one feature vector for the second encounter, and three feature vectors for the third encounter. The feature matrix is illustrated in Supplementary Fig. 2.

2.4. Predictive model construction

LSTM networks are a modified version of recurrent neural networks capable of learning long-term dependencies. We employed the LSTM model in our problem because compared to genetic RNNs, LSTM networks are better suited to maintain past information, and they solve the vanishing/exploding gradient problem which leads to instability when training RNN models with long sequences [24]. Based on our experiment with different structures of the LSTM model on different layers (2/3/4/5), and different dimensions of each layer (32/64/128/256/512), we chose the one of best performance as the proposed LSTM model. We implemented an LSTM based network composed of two LSTM layers of 64 units. A sigmoid function was applied in the last layer. We determined the parameters using the Adam optimizer with binary cross-entropy loss function. Furthermore, to better differentiate the importance of features, we applied the attention mechanism [25] to the LSTM model for comparison. Attention is a mechanism for a neural network to have the ability to focus on a subset of its inputs (or features) by transferring input vectors into weighted embeddings. The structure of the attention model is shown in Supplementary Fig. 2.

2.5. Baseline methods

Besides the proposed LSTM based models, we included other common methods for comparison, including decision tree, random forest, logistic regression and DNN. The feature matrix is flattened to a single vector as the input to those models. Logistic regression is implemented with default settings of L2 penalty, c value of 1.0, limited-memory BFGS (lbfgs) solver. Random forest has a number of trees is 100, criterion function is GINI impurity and no limitation of max depth of single tree. Decision tree also took the GINI impurity criterion function, but the max depth of each tree is 500. The DNN model is composed of 4 dense layers, each of first 3 layers has a dimension of 512, each is followed by a dropout of 0.3, and the last layer has a dimension of 8, then connected to a sigmoid layer as output. It has the same optimizer setting as the LSTM model whose learning rate is 0.01, optimizer is Adam, loss function is

binary cross-entropy. Additional explanations can be found in Supplementary Table 3.

In the experiment, the software implementation environment is Python (2.7). The Python Scikit-Learn package [26]was used for classical machine learning methods, whereas Python Tensorflow [27] and Python Keras [28] were used for deep learning. Supporting libraries include Python Numpy [29] and Python Pandas [30]. For training, we used an NVIDIA Tesla V100 with 16 GB memory.

2.6. Ranking of features

To help researchers or clinicians understand the prediction, we want to explain how different features impact each model's result. Due to the large number of features and the complexity of their structures, interpreting deep learning is a challenging task and has been an active research area [31]. . W.We used the permutation importance method to assess the importance of each feature in our deep learning models. This method measures the loss of performance by randomizing the values of a feature through shuffling [32]. We prioritized the AUROC score for comparison since it measures the probability changes for all patient instances other than a binary result, better for comparison of the overall prediction performance. Using the model of best performance based on F-1 and AUROC score, we would generate the top 50 important features. To implement permutation importance, we randomly shuffled the values for each feature one at a time in the test set, and then fed them into the model. After that, we measured the decrease for this feature. We repeated the process 10 times to calculate the average AUROC score decrease of a feature as the importance measurement.

3. Data availability statement

Health Facts database is a commercial database provided by Cerner, and the use of the database is based on an agreement between Stony Brook University Hospital and Cerner, and sharing of the dataset extracted from Cerner is prohibited. Source codes are available at https://github.com/StonyBrookDB/odprediction.

4. Results

4.1. Prediction performance

To comprehensively evaluate the performance, we calculated all important metrics including precision, recall, F-1 score, area under receiver operating characteristic curve (AUCROC) score and area under precision recall curve for each algorithm. We have repeated each experiment 30 times for each method, and the average scores and standard deviations are reported in Table 2. Fig. 1 shows the ROC curves for all methods.

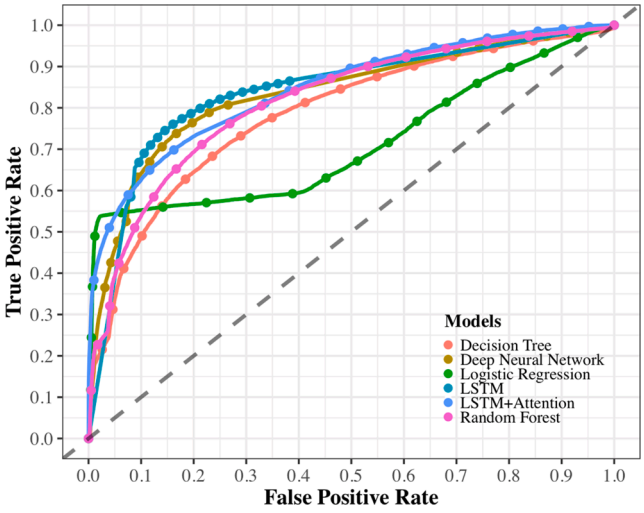

Fig. 1. ROC curves for all five methods.

The experiments demonstrate promising results for predicting opioid poisoning with our LSTM based model. LSTM with attention mechanism (LSTM + Attention) achieved an F-1 score of 0.7815, the highest precision (0.8128) and the highest AUC score (0.8449) among all models. The prototypical LSTM also has better performance than other methods in recall, F-1 score and AUC score, and achieved the best recall (0.7616). Furtherly, to prove that our LSTM model outperforms the DNN model, we applied *t*-test to compare the result between the two methods. The result shows that LSTM is considered to be significantly better than DNN model (at Supplementary Table 5).

4.2. Handling missing values

Missing values such as missing lab tests are common in EHR. For some lab tests, there is a large difference in the frequency of missing values between the two classes of patients. In the case of differential atypical lymph, 20.06% of values for negative patients are missing, while 54.31% are missing for positive patients. That difference may affect prediction performance. To avoid this problem, we performed a sensitivity analysis to compare the results before and after removing the lab test features with a high proportion of missing values. We removed all the lab tests with more than 5% difference in missing value proportion (Supplementary Table 4). Since the change in performance was minimal, we concluded that these features were nonessential.

4.3. Top feature analysis

Based on the method of best performance, LSTM + Attention, we applied the permutation method to generate a list of 50 features ranked by the decrease value of ROC AUC score. They are shown in Table 3 and Fig. 2. Since Body Mass Index (BMI) is calculated from height and weight, we combined them as a single feature. We also combined systolic blood pressure and diastolic blood pressure.

Among the list, there are 12 features on medications, including the aggregated feature MME, 8 clinical events, 5 diagnoses, and 25 laboratory tests. Among the top 10 features in the list, medications dominate, followed by clinical events and lab tests. "N02A Opioids" represents the quantity of opioid medications in ATC Level 3 code N02A, comprising strong analgesics of the opiate type and analgesics with similar structure or action, and ranks to the top of the list. "Pain Scale Score" implies the need for pain management, where opioid is commonly prescribed. "A07D: Antipropulsives" comprises agents which reduce gastrointestinal motility, e.g., diphenoxylate (opioid medication used in a combination drug with atropine for the treatment of diarrhea), and loperamide (active ingredient of Imodium, an opioid receptor agonist that can be used to treat diarrhea) whose over the counter use has been restricted by the FDA due to significant non-medical use and opioid overdose. "N01A: Anesthetics, general" comprises agents which produce general anesthesia, surgical analgesia or neuroleptanalgesia, including opioid anesthetics (N01AH). "Alcohol Use" is among top co-occurrence diagnoses with opioid overdose [33]. Hypotension is commonly associated with opioid intoxication [34]. Lab test "Mean Corpuscular Hemoglobin" (MCH) level can have a significant increase in heroin and opium withdrawal groups [35] and "Red Blood Cell Distribution Width (RDW)" can be much higher in heroin use patients [36]. Not surprisingly, MME comes right after RDW as another significant feature.

The permutation based feature ranking has its own limitations, such as dependency between features. "N02B: Other analgesics and antipyretics", mostly non-opioid medications, can often be combined with opioid analgesics "N02A: Opioids" when managing moderate to severe pain [37]. To understand the correlation among the top features, we generated a heatmap of feature correlations using pearson correlation coefficient, as shown in Supplementary Figure 3. Besides, we have listed

Table 3
Top 50 important features for OD prediction.

Rank	Category	Description	Importance Measurement (AUROC score decrease)
1	Medication	N02A: Opioids	0.070492
2	Clinical Event	Pain Scale Score	0.061711
3	Medication	A07D: Antipropulsives	0.054692
4	Medication	N01A: Anesthetics, general	0.051445
5	Clinical Event	Alcohol Use	0.040293
6	Medication	N02B: Other analgesics and antipyretics	0.032074
7	Clinical Event	Blood Pressure	0.019784
8	Medication	N05C	0.018649
9	Laboratory Test	Mean Corpuscular Hemoglobin	0.018202
10	Laboratory Test	Red Blood Cell Distribution Width (RDW)	0.01609
11	Medication	MME	0.010222
12	Clinical Event	Smoke, Exposure to Tobacco Smoke	0.010002
13	Medication	G02A: Uterotonics	0.009293
14	Laboratory Test	Blood Urea Nitrogen	0.008829
15	Medication	R05D: Cough and cold preparations	0.008315
16	Laboratory Test	Alkaline Phosphatase, Serum	0.007975
17	Clinical Event	Heart Rate	0.006447
18	Laboratory Test	Chloride, Serum	0.006336
19	Clinical Event	Height Weight BMI	0.006055
20	Laboratory Test	Monocyte Count	0.004942
21	Medication	R02A: Throat preparations	0.004664
22	Clinical Event	Tobacco Use	0.004457
23	Laboratory Test	Albumin, Serum	0.004322
24	Laboratory Test	Lymphocyte Absolute Count	0.003803
25	Laboratory Test	Basophils Percent	0.003413
26	Diagnosis	Other and unspecified disorders of back	0.003364
27	Laboratory Test	Hemoglobin	0.00313
28	Diagnosis	Nondependent abuse of drugs	0.003078
29	Laboratory Test	Red Blood Cell Count	0.003074
30	Laboratory Test	Glomerular Filtration Rate Estimated	0.003068
31	Laboratory Test	Blood Urea Nitrogen	0.00305
32	Medication	H01B: Posterior pituitary lobe hormones	0.002934
33	Clinical Event	BSA, Body Surface Area	0.002931
34	Laboratory Test	Potassium, Serum	0.002827
35	Laboratory Test	Hematocrit	0.00267
36	Diagnosis	Pain, not elsewhere classified	0.002598
37	Laboratory Test	Calcium, Serum	0.002467
38	Laboratory Test	Creatinine, Serum Quantitative	0.00239
39	Medication	N05B: Anxiolytics	0.00234
40	Laboratory Test	Blood Gas CO2 Total, Arterial	0.002286
41	Laboratory Test	Lymphocyte Percent	0.002279
42	Laboratory Test	Drug dependence	0.002221
43	Laboratory Test	UA White Blood Cell	0.002216
44	Medication	B05C: Irrigating solutions	0.002009
45	Diagnosis	Essential (Primary) Hypertension	0.001985
46	Laboratory Test	Neutrophil Percent	0.001952
47	Laboratory Test	Erythrocytes Blood Automated Count	0.001867
48	Diagnosis	Osteoarthrosis and allied disorders	0.001823
49	Laboratory Test	Prothrombin Time	0.001733
50	Laboratory Test	Mean Platelet Volume	0.001693

separately the correlation heatmap between only lab test features in Supplementary Figure 4. As lab tests with missing values are imputed with median values, they tend to have a higher correlation than between other types of features.

All feature pairs with a correlation higher than 0.5 are shown in Supplementary Table 6. For example, "N02B: Other analgesics and antipyretics" and "N02A: Opioids" have a high correlation coefficient of 0.671. We also found that many lab tests have high correlation between each other. We conducted a study to evaluate the performance of our model by removing all lab tests, and the results were F-1 score at 0.7654, recall at 0.7243, precision at 0.7729 and AUROC at 0.8318. Compared to the results including lab tests with F-1 score at 0.7815, precision at 0.8128, recall at 0.7512, AUROC at 0.8449, we found that lab tests features only had a marginal improvement on the performance, e.g, an increase of 0.0161 for F-1 score and an increase of 0.0131 of AUROC. While there are many lab tests based features appearing in the top 50 list, due to the correlation among lab tests, largely due to median imputation for missing values, the predictive power of lab tests is limited. We will continue to explore emerging interpretable machine learning techniques to help to better understand the prediction in our future work.

5. Discussion

The CDC's strategy on combating the opioid epidemic has a major focus on better data and tools for evidence-based decision-making [38]. HHS' five point strategy emphasizes actionable data that can be used to target interventions [39]. The availability of large EHR datasets, coinciding with developments in predictive modeling, creates an unparalleled opportunity for early, targeted interventions to prevent opioid overdose. Sequential deep learning models have better modeling of disease progression, and thus can be an ideal framework for building such predictions. We developed sequential models using LSTM for opioid overdose prediction, and achieved promising results.

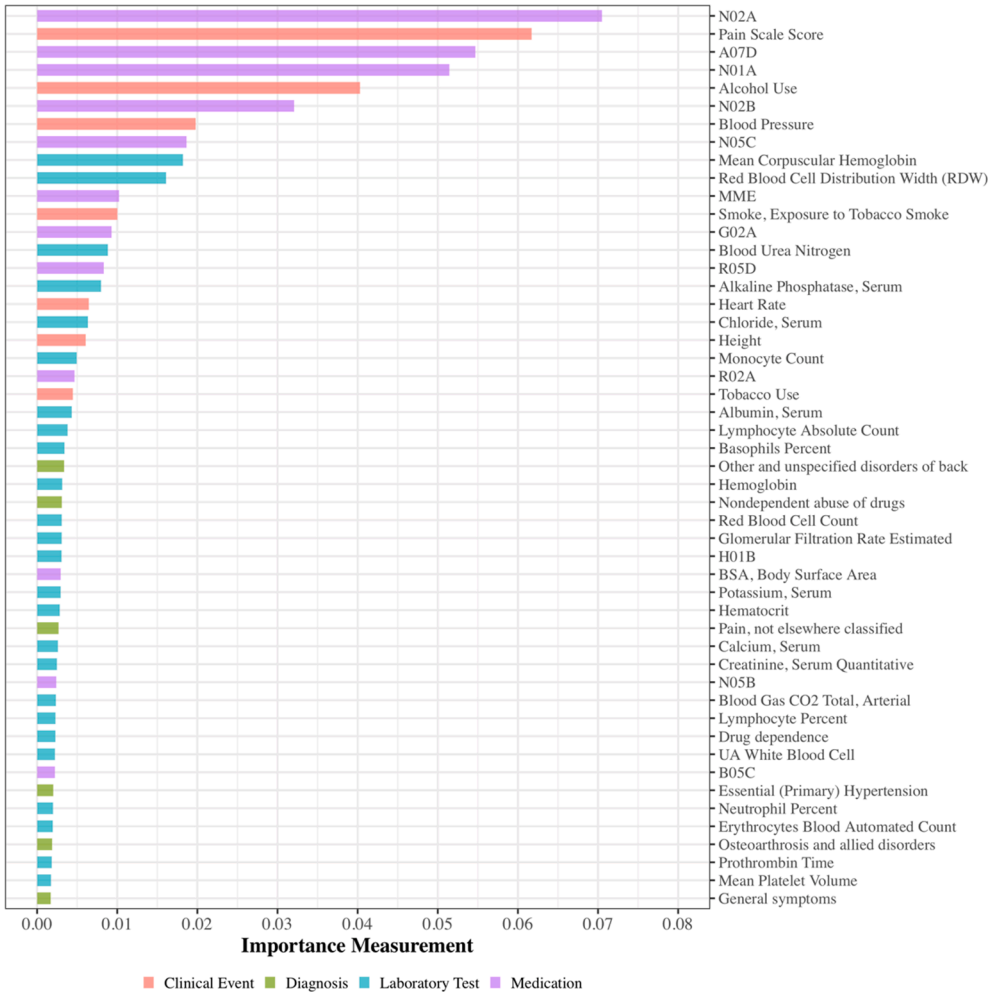

Fig. 2. Top 50 important features for OD prediction with LSTM.

5.1. Comparison with related work

We have identified multiple previous studies working on opioid related diseases prediction. Lo-Ciganic (2018) applied dense neural network (DNN) models on 268 hand-crafted clinical features to predict opioid overdose risk within 3 months after opioid prescription [3]. They achieve a highest recall of 0.92 and AUCROC of 0.9, but precision is quite low, and area under precision recall curve is only 0.2–0.36. That means they had an outstanding ability to identify overdose patients, but they will misclassify many negative patients as overdose patients. Hasan et al. (2019) proposed an analytic framework using multiple machine learning feature selection techniques and algorithms to predict the risk of opioid overdose [40]. Prieto and colleagues (2020) collected paramedic trip report documents with related keywords like “naloxone” and/or “heroin” and applied machine learning models to identify non-medical opioid use [41]. Ellis RJ and colleagues (2019) employed statistical methods including Gini importance and Wilcoxon rank-sum tests to extract the most relevant features, and used a random forest classifier to predict ICD-9 opioid dependence [42], whose highest F1 score is 0.776. Che et al [8] used a recurrent neural network (RNN) on diagnosis, procedure and prescription information to classify opioid users into long term users, short term users, and opioid dependent patients. Their best AUCROC score for predicting opioid dependent patients who can be identified as OD is around 0.8.

Compared to previous work, our study has overcome their limitations. For feature engineering, most previous works are based on clinical knowledge from domain experts, which are cumbersome and could be incomplete. Our model takes advantage of the deep learning model’s ability to process large scales of data without relying on domain knowledge. Thus, more clinical features can be included and even unexpected relationships can be discovered. Another advantage of the LSTM model employed in our methodology is its enhanced ability to remember knowledge from long sequences, while being resilient to the vanishing gradient problem [24].

There is a tradeoff between precision and recall. Precision value can be increased by revising the parameters of the model while decreasing recall value, and vice versa. F-1 score, as a weighted average of them, is a more informative metric to evaluate the prediction performance. Thus in this study, we used the F-1 score as a major metric to support the model development and evaluation.

6. Limitations

This study has a few limitations. First, since the population is derived from patients that were prescribed opioids, the results might not generalize well to patients who used non-prescribed opioids. Secondly, this study captured only opioid overdose diagnoses in electronic health records, potentially missing patients with unrecorded overdose, and patients who died from overdose at home.

7. Future work

Our model can benefit from integrating unstructured information. Natural language processing algorithms have been extremely successful

in processing clinical notes [9–11] to explore additional knowledge not available in structured EHR data. One current limitation is class imbalance in the prediction, since there are more negative cases than positive cases. To improve the prediction performance, one potential solution is to use generative adversarial networks [43] to generate more examples representing positive patients. Transparency and explainability of the model's decision are vital for clinical decision support. To deliver more interpretable results for clinical decision support, we will explore deep learning visualization tools to understand the model interactively for understanding the decision process and critical risk factors.

8. Conclusions

The opioid epidemic is already a severe public health problem in the United States, which requires critical knowledge and solutions to combat. Machine learning based predictive models can serve as tools for clinical decision support to prevent development of opioid overdose. Our work shows that sequential deep learning models are promising for opioid overdose risk detection. Our methodology can continue to be improved with more advanced deep learning techniques.

9. Ethics declaration

All the data used in the study are de-identified data, and the study is approved by Stony Brook University.

Author contributions

FW, XD, JD, WH, SR, RNR, MS and JH conceived and designed the study. XD, JD, WH, SR, RNR and FW contributed to acquisition, analysis, or interpretation of data. XD drafted the manuscript. All authors contributed to critical revision. XD and WH contributed to statistical analysis. FW obtained the funding and administrated the project. XD performed verification of the underlying data. All authors read and approved the final manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was funded partially by the Stony Brook University OVPR Seed Grant 1158484-63845-6. We thank Kayley Abell-Hart for English language editing, and thank Janos Hajagos for data access support.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103725.

References

[1] R.A. Rudd, P. Seth, F. David, L. Scholl, Increases in drug and opioid-involved overdose deaths - United States, 2010–2015, MMWR Morb. Mortal Wkly Rep. 65 (5051) (2016) 1445–1452.
[2] C. Florence, F. Luo, L. Xu, C.J.M.C. Zhou, The economic burden of prescription opioid overdose, abuse and dependence in the United States, 2013, 54(10) (2016) 901.
[3] W.-H. Lo-Ciganic, J.L. Huang, H.H. Zhang, et al., Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions, J. JAMA network open 2 (3) (2019) 190968.
[4] X. Dong, S. Rashidian, Y. Wang, et al., Machine learning based opioid overdose prediction using electronic health records, in: AMIA Annual Symposium Proceedings 2019, 2019, p. 389.
[5] Rashidian S, Hajagos J, Moffitt R, Wang F, Dong X, Abell-Hart K, Noel K, Gupta R, Tharakan M, Lingam V, Saltz J. Disease phenotyping using deep learning: A diabetes case study. arXiv preprint arXiv:1811.11818. 2018 Nov 28.
[6] S. Rashidian, X. Dong, A. Avadhani, P. Poddar, F. Wang, Effective scalable and integrative geocoding for massive address datasets, in: Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2017, pp. 1–10.
[7] S. Rashidian, J. Hajagos, R.A. Moffitt, et al., Deep learning on electronic health records to improve disease coding accuracy, in: AMIA Summits on Translational Science Proceedings, 2019, 2019, p. 620.
[8] Z. Che, J.S. Sauver, H. Liu, Y. Liu, Deep learning solutions for classifying patients on opioid use, in: AMIA Annual Symposium Proceedings 2017, 2017, p. 525.
[9] Liu J, Zhang Z, Razavian N. Deep ehr: Chronic disease prediction using medical notes. InMachine Learning for Healthcare Conference 2018 Nov 29 (pp. 440-464). PMLR.
[10] A.N. Jagannatha, H. Yu, Structured prediction models for RNN based sequence labeling in clinical text, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing 2016, 2016, p. 856.
[11] Edward Choi, et al., Doctor AI: Predicting Clinical Events via Recurrent Neural Networks, in: JMLR workshop and conference proceedings 56, 2016, pp. 301–318.
[12] Choi E, Schuetz A, Stewart WF, Sun J. Medical concept representation learning from electronic health records and its application on heart failure prediction. arXiv preprint arXiv:1602.03686. 2016 Feb 11.
[13] University of Texas Health Science Center at Houston. SBMI Data Service. https://sbmi.uth.edu/sbmi-data-service/data-set/cerner/ (accessed March 21, 2020).
[14] D.S. Wishart, Y.D. Feunang, A.C. Guo, et al., DrugBank 5.0: a major update to the DrugBank database for 2018 46(D1) (2018) 1074–1082.
[15] The New Zealand Ministry of Health. Mapping between ICD-10 and ICD-9. https://www.health.govt.nz/nz-health-statistics/data-references/mapping-tools/mapping-between-icd-10-and-icd-9 (accessed March 21, 2021).
[16] R.K. Portenoy, K.M.J.P. Foley, Chronic use of opioid analgesics in non-malignant pain: report of 38 cases 25(2) (1986) 171–186.
[17] Centers for Disease Control and Prevention. SCREENING LIST OF ICD-9-CM CODES FOR CASEFINDING. https://www.cdc.gov/cancer/apps/ccr/icd9cm_codes.pdf (accessed March 21, 2021).
[18] Centers for Disease Control and Prevention. ICD-10-CM Table of NEOPLASMS. https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Publications/ICD10CM/2019/icd10cm_neoplasm_2019.pdf (accessed March 21, 2021).
[19] G. Miller, H. Britt, A new drug classification for computer systems: the ATC extension code, J. Int. J. Bio-medical Comput. 40 (2) (1995) 121–124.
[20] The Centers for Medicare & Medicaid Services. Opioid oral morphine milligram equivalent (MME) conversion factors. https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovContra/Downloads/Oral-MME-CFs-vFeb-2018.pdf (accessed March 21, 2021).
[21] Centers for Disease Control and Prevention. Calculating total daily dose of opioids for safer dosage. https://www.cdc.gov/drugoverdose/pdf/calculating_total_daily_dose-a.pdf (accessed March 21, 2021).
[22] T.R. Frieden, D. Houry, Reducing the risks of relief—the CDC opioid-prescribing guideline, New England J. Med. 374 (16) (2016) 1501–1504.
[23] K. Potdar, T.S. Pardawala, CDJIjoca Pai, A comparative study of categorical variable encoding techniques for neural network classifiers 175(4) (2017) 7–9.
[24] S. Hochreiter, J. Schmidhuber, Long short-term memory, J. Neural Comput. 9 (8) (1997) 1735–1780.
[25] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, arXiv preprint arXiv:1409.0473, 2014 Sep 1.
[26] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., Scikit-learn: machine learning in, Python. 12 (2011) 2825–2830.
[27] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, Tensorflow: A system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), 2016, pp. 265–283.
[28] A. Gulli, S. Pal, Deep learning with Keras, Packt Publishing Ltd, 2017 Apr 26.
[29] E. Bressert, SciPy and NumPy: an overview for developers, O'Reilly Media, Inc., 2012 Nov 15.
[30] W. McKinney, Pandas: a foundational Python library for data analysis and statistics, Python for High Performance and Scientific Computing 14(9) (2011).
[31] Q.S. Zhang, S.C. Zhu, Visual interpretability for deep learning: a survey, Front. Inform. Technol. Electron. Eng. 19 (1) (2018) 27–39.
[32] L. Breiman, Random forests, J. Machine Learning 45 (1) (2001) 5–32.
[33] A.H. Rogers, J.M. Shepherd, D.J. Paulus, M.F. Orr, J.W. Ditre, J. Bakhshaie, M. J. Zvolensky, The interaction of alcohol use and cannabis use problems in relation to opioid misuse among adults with chronic pain, International journal of behavioral medicine 26 (5) (2019 Oct 1) 569–575.
[34] A. Fareed, S. Stout, J. Casarella, S. Vayalapalli, J. Cox, K. Drexler, Illicit opioid intoxication: diagnosis and treatment. Substance abuse: research and treatment, 2011 Jan;5:SART-7090.
[35] T. Haghpanah, M. Afarinesh, K. Divsalar, A review on hematological factors in opioid-dependent people (opium and heroin) after the withdrawal period, Addiction Health 2 (1–2) (2010) 9.
[36] D. Guzel, A.B. Yazici, E. Yazici, A. Erol, Evaluation of immunomodulatory and hematologic cell outcome in heroin/opioid addicts, Journal of addiction (2018 Dec 9).
[37] D.E. Becker, J.C. Phero, Drug therapy in dental practice: nonopioid and opioid analgesics, Anesthesia Progr. 52 (4) (2005) 140–149.

[38] Centers for Disease Control and Prevention. Understanding the Epidemic. https://www.cdc.gov/drugoverdose/epidemic/index.html (accessed March 21, 2021).
[39] The U.S. Department of Health and Human Services. Strategy to Combat Opioid Abuse, Misuse, and Overdose. https://www.hhs.gov/opioids/sites/default/files/2018-09/opioid-fivepoint-strategy-20180917-508compliant.pdf (accessed March 21, 2021).
[40] M.M. Hasan, M.R. Patel, A.S. Modestino, et al., A Novel Big Data Analytics Framework to Predict the Risk of Opioid Use Disorder, arXiv preprint arXiv: 1904.03524 (accessed Apr 6, 2019).
[41] J.T. Prieto, K. Scott, D. McEwen, et al., The detection of opioid misuse and heroin use from paramedic response documentation: machine learning for improved surveillance 22(1) (2020) 15645.
[42] R.J. Ellis, Z. Wang, N. Genes, A. Ma'ayan, Predicting opioid dependence from electronic health records with machine learning, J. BioData Mining 12 (1) (2019) 3.
[43] M. Rezaei, H. Yang, C. Meinel, Generative adversarial framework for learning multiple clinical tasks. Digital Image Computing: Techniques and Applications (DICTA) 2018 Dec 10, pp. 1–8.