Supplementary Content

eAppendix. Machine learning algorithms

eTable 1. Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

eTable 2. Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

eTable 3. Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the chi^2 test of independence were <0.001 unless otherwise indicated.

eTable 4. Characteristics of study participants between training and validation groups using 2017 data.

eTable 5. Candidate predictors used to train ML algorithms.

eTable 6. Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms. Training and validation were done using 2017 data (n=393,979); another independent validation was performed using 2018 data (n=393,023).

eFigure1. Schematic of study design and feature generation

eFigure2. Feature importance from logistic regression and tree-based (XGBoost) classifiers using the 2018 validation set.

eFigure3. Shapley values and feature impact in the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome.

eFigure 4. Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression (L1) classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.

eReferences.

Sharma V, *et al. BMJ Open* 2021; 11:e043964. doi: 10.1136/bmjopen-2020-043964

eAppendix. Machine Learning Algorithms

Introduction

While there are always updates and new methods coming up in the fields of machine learning, in this study, we have focused on some of the most reliable and proven approaches for predictive modelling which are explainable and popularly used in previous studies of similar nature.

Logistic Regression

Regression analysis models the relationship between a dependent variable and a set of independent variables [1]. Typically, this includes understanding how the value of the dependent variable changes with the changes in the values of independent variables. Logistic regression [1] uses the logistic function to model a binary dependent variable, where, based on the values of the independent variables the model can approximate one of the two classes, the instance belongs to. This basic binary model can be extended to deal with multiple classes (e.g. One-vs-all classifiers). However, logistic regression is only capable of modeling a linear relationship of independent variables to the dependent variable, hence limited to problems with linear decision boundaries. We used the sci-kit learn library in our experiments[6] and found L1 regularization to be more effective.

Ridge Classifier

We used the ridge classifier implemented in the Scikit learn library[5]. It implements a classifier using ridge regression which uses an L2 regularization on the least square objective function. The library converts the labels into -1 and 1 and fits a linear regression on the converted labels with the regularization.

Random Forest

Random forest is a tree ensemble learning algorithm that has wide applicability in many domains[1]. Random forest is a nonlinear learning algorithm, which arrives at nonlinear decision boundaries by independently combining multiple decision trees. Each individual decision tree in the forest can be grown independently of each other on a subset of the training data. Random forests are mainly sensitive to the number of trees, the depth of a tree and the number of covariates randomly chosen to split at each node[1]. These hyper-parameters can be tuned to find the best configuration of every dataset. Random Forests, in general, are less prone to overfit since they always grow individual trees on a subset of the training data[1]. At prediction time, the decision of each tree is aggregated to compute the final prediction.

Neural Networks (NN)

Neural networks are another collection of non-linear learning algorithms with high representation power. They are known to be able to find mappings from an input to an output from a larger non-linear function space [2]. This ability to represent a larger space of nonlinear

functions has shown to be very effective recently in many application domains such as natural language processing, computer vision, genomics, computer games and health[2]. Neural networks come in many flavors learning nonlinear mapping of different types of data such as Convolutional NNs being most effective with images and Recurrent NNs for time series and language data. Identifying the most effective neural network structure is one of the difficult and the most time-consuming aspect of applying neural networks to new application domains and data. Generally, neural networks try to exploit the relationships in the raw unstructured data (eg: image and text) presented to the network but with more structured data such as health records and ICD codes learning relationships is much complex. Our neural network models are mainly based on densely connected hidden layers with ReLu[6] activation function. We used the cross-entropy loss for the binary classification Adam optimizer. We used a simple feed forward network using Sklearn MLP classifier with hyperparameter tuning for the NN.

Boosted Learning Algorithms

Boosting is a process to ensemble multiple base learning algorithms to arrive at better overall performance than any individual base learner[1]. In contrast to independently building multiple models from the subsets of the data, boosting re-weights the training data every time a model is learned for future models. This weighting happens to give more preference to currently misclassified data points in the next round compared to the correctly classified data points. Therefore future learners try to do better on the misclassified data points leading to a collection base learners having a better-combined prediction. This process is sequential so each base learner is dependent on the output of the previously trained model (it is worthy to note XGBoost provides a parallel tree boosting alternative). In our work, we have experimented with several boosting meta-learning algorithms such as XGBoost[7], AdaBoost[5] and GBM[5]. XGBoost uses a variant of trees as the base learner whereas AdaBoost (from Sci-kit learn) can use many ML algorithms as base learners. GBM uses logistic regression by default as the base learner. We used all 3 types of boosting with tuned hyperparameters for comparison.

Naive Bayes

Naive Bayes is based on the Bayes theorem with a strong independence assumption between the covariates[1]. This assumption helps in building a simple probabilistic model for learning and inference. Naive Bayes coefficients scale linearly with the number of covariates making this a suitable model for high-dimensional data. We used Naive Bayes as a simple baseline learning algorithm for comparison.

Support Vector Machines (SVM)

SVMs[4] are maximum margin classifiers optimizing for learning a hyperplane having the maximum distance away from each of the class data points[1]. SVM is a linear classifier but with the kernel trick to map the inputs to the higher dimensional space, it can learn nonlinear decision boundaries in the input space. SVMs are very effective binary classifiers with the kernel trick[1]. With larger datasets, SVMs tend to become more computationally intensive.

eTable 1. Diagnostic codes used to exclude patients who had cancer, were pregnant, or were under palliative care.

Condition	**ICD 9**	**ICD 10**
Cancer	140.x - 239.x	C00.x - C99.x, D00.x - D49.x
Pregnancy	630.x - 679.x	O00.x - O99.x
Palliative	V66	Z51.0, Z51.1, Z51.5

eTable 2. Diagnostic codes used to identify the defined study outcome from emergency visit, hospitalization and death data.

ICD 10	**Condition**
T40.x	Poisoning by, adverse effect of and underdosing of narcotics and psychodysleptics
F55.x	Abuse of non-psychoactive substances
F11.x - F19.x	Mental and behavioral disorders due to psychoactive substance use

Sharma V, *et al. BMJ Open* 2021; 11:e043964. doi: 10.1136/bmjopen-2020-043964

eTable 3. Baseline characteristics of study patients (n=392,979). Co-morbidities were determined using Elixhauser criteria. All p-values in the chi^2 test of independence were <0.001 unless otherwise indicated.

Characteristic	Number without Event n=386,371	Percent	Number with Event n=6,608	Percent
Age:				
Mean (SD)	48.1 (16.4)	--	41.2 (12.4)	--
18-45	162057	41.9	3466	52.4
45-65	154632	40.0	2656	40.2
>65*	69682	18.0	486	7.4
Male	197491	50.3	3922	59.4
Female	194794	49.7	2686	40.6
Alcohol Disorder	66320	16.9	5220	79.0
Arrhythmia	90621	23.1	1959	29.6
Blood Loss Anemia	1164	0.3	82	1.2
Congestive Heart Failure	18954	4.8	565	8.6
Coagulopathy	8053	2.1	356	5.4
Deficiency Anemia	34188	8.7	971	14.7
Depression	159140	40.6	5518	83.5
Diabetes**	64132	16.3	1408	21.3
Substance Abuse Disorder	74678	19.0	5485	83.0
Fluid Disorder	42690	10.9	3012	45.6
Hypertension**	140171	35.7	2624	39.7
Hypothyroidism	45519	11.6	601	9.1
Injury^	195688	49.9	5541	83.9
Liver Disorder	21656	5.5	1588	24.0
Neurologic Disorder	230490	58.8	5387	81.5
Obesity	63393	16.2	970	14.7
Poisoning^	17434	4.4	2775	42.0
Psychoses	35870	9.1	3162	47.9
Renal Disorder	16166	4.1	499	7.6
Rheumatoid Conditions	111458	28.4	3157	47.8
HIV Infection	1098	0.3	141	2.1
Paralysis	3874	1.0	187	2.8
Peptic Ulcer Disease	11728	3.0	509	7.7
Pulmonary Circulation Disorder	9611	2.4	430	6.5
Chronic Pulmonary Disease	102990	26.3	2913	44.1
Peripheral Vascular Disease	14467	3.7	389	5.9
Valvular Disease	7308	1.9	226	3.4
Weight Loss	16207	4.1	747	11.3

*p-value for age >65 is an estimated 0.037

^ Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50

** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each

eTable 4. Characteristics of study participants between training and validation groups using 2017 data.

Characteristic	Number in training group N=275,150~	Percent	Number in validation group N=117,829~	Percent
Age:				
Mean (SD)	48.3 (16)	--	48.2 (16)	--
18-45	114356	41.5	49909	42.3
45-65	111859	40.7	47132	40.0
>65	48935	17.8	20788	17.6
Male	138603	48.5	59339	48.4
Female	136545	47.8	58490	47.7
Alcohol Disorder	46792	16.4	20199	16.5
Arrhythmia	63637	22.3	27201	22.2
Blood Loss Anemia	839	0.3	336	0.3
Congestive Heart Failure	13320	4.7	5694	4.6
Coagulopathy	5697	2.0	2393	2.0
Deficiency Anemia	24096	8.4	10179	8.3
Depression	112080	39.2	47628	38.9
Diabetes**	45131	15.8	19144	15.6
Substance Abuse Disorder	52609	18.4	22713	18.5
Fluid Disorder	30272	10.6	12780	10.4
Hypertension**	98546	34.5	41840	34.1
Hypothyroidism	31908	11.2	13666	11.2
Injury*	137423	48.1	58865	48.0
Liver Disorder	15252	5.3	6567	5.4
Neurologic Disorder	161706	56.5	69341	56.6
Obesity	44607	15.6	18882	15.4
Poisoning*	12503	4.4	5293	4.3
Psychoses	25422	8.9	10860	8.9
Renal Disorder	11403	4.0	4817	3.9
Rheumatoid Conditions	78268	27.4	33420	27.3
HIV Infection	774	0.3	336	0.3
Paralysis	2717	1.0	1176	1.0
Peptic Ulcer Disease	8239	2.9	3533	2.9
Pulmonary Circulation Disorder	6771	2.4	2877	2.3
Chronic Pulmonary Disease	72265	25.3	30949	25.3

Peripheral Vascular Disease	10228	3.6	4278	3.5
Valvular Disease	5111	1.8	2215	1.8
Weight Loss	11477	4.0	4790	3.9

Note: p-values for chi^2 test of independence were all >0.06 when comparing training and validation sets.
*Injury: ICD10: S00-T98; Poisoning: ICD10: T36-T50
** Complicated, uncomplicated diabetes and hypertension were collapsed into one category each

eTable 5. Anatomical Therapeutic Chemical classification of opioid molecules used for this study and candidate predictors used to train ML algorithms.

Category (data source)	**Description**
ATC codes used to identify opioids from PIN data	N01AH01, N01AH03, N01AH06, N07BC01, N07BC02, N07BC51, R05DA03, R05DA04, R05DA09, R05DA20, N02A
Opioid molecules used in this study	alfentanil, butorphanol, codeine, diamorphine, fentanyl, hydrocodone, hydromorphone, meperidine, morphine, oxycodone, oxymorphone, pentazocine, sufentanil, tapentadol, tramadol
Demographic information (PIN)	age, sex, postal codes, mean income
Drug utilization history (PIN)	drug dispenses in past 30 days using on ATC codes, oral morphine equivalents, concurrent use with benzodiazepines defined as at least 7 days of cumulative concurrent use in the 30 days prior to dispensation, number of dispensations and unique molecules of opioids and benzodiazepines
Health care utilization (PIN DAD)	flags for previous hospitalizations and emergency department visits, number of unique providers
ICD based co-morbidities (DAD, NACRS, Claims)	Elixhauser condition flags based on the past 5 years of claims, hospitalizations, and emergency visits.

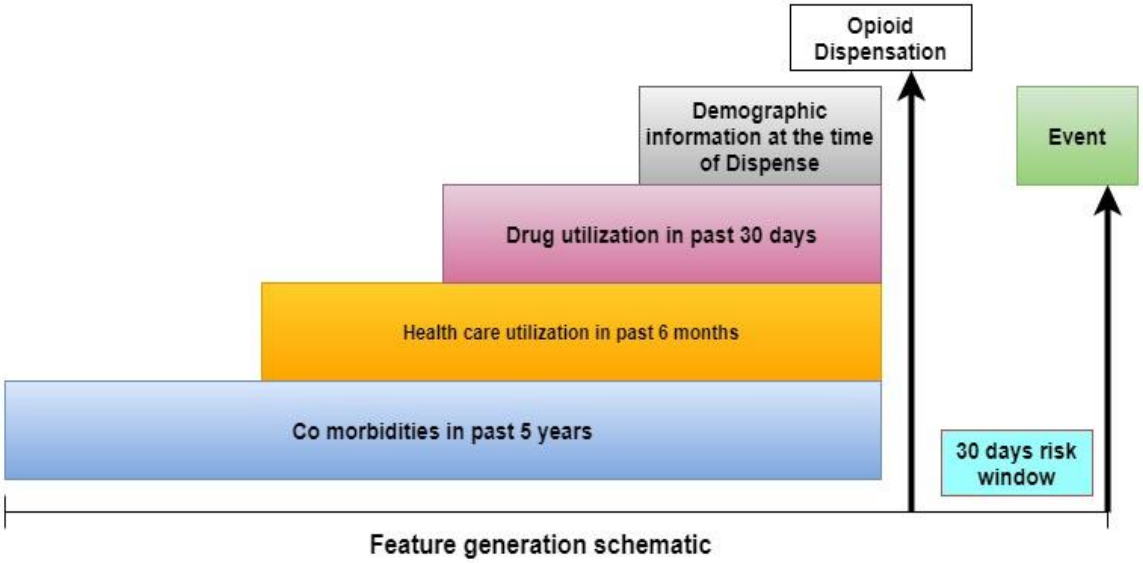
Note: ATC- Anatomical Therapeutic Chemical classification (https://www.whocc.no/atc_ddd_index); PIN- Pharmaceutical Information Network; ICD- International Statistical Classification of Diseases and Related Health Problems, World Health Organization; total number of features 283

eTable 6. Discrimination performance using area under the receiver operating characteristic curve (AUROC) of various ML algorithms using all features (demographics, health utilization, prescription history, co-morbidities). Training and validation were done using 2017 data (n=393,979); another independent validation was performed using 2018 data (n=393,023).

Algorithm	**Train**	**Validation 2017**	**Validation 2018**
XGBoost Classifier	0.897	0.870	0.884
Logistic Regression	0.887	0.869	0.884
Gradient Boosting Classifier	0.898	0.868	0.883
AdaBoost Classifier	0.884	0.868	0.882
Random Forest Classifier	0.909	0.863	0.881
Ridge Classifier	0.895	0.863	0.879
SVM	0.896	0.860	0.878
Gaussian Naive Bayes	0.846	0.826	0.847
Decision Tree Classifier	0.919	0.791	0.822
Neural Networks	0.827	0.804	0.821

Note: Logistic regression used L1 (lasso) parameter regularization

eFigure 1. Schematic of study design and feature generation



eFigure2. Feature importance from logistic regression and tree-based XGBoost classifiers using the 2018 validation set. The logistic regression classifier relied more on co-morbidity data from DAD, NACRS, and Claims databases; XGBoost classifier relied more on data from the PIN database. AUROCs for both classifiers were similar at 0.88.

Logistic Regression		XGBoost	
history of drug abuse	1.00	age at dispensation	1.00
age at dispensation	0.65	number of prescriptions dispensed in previous 30 days	1.00
history of prior hospitalization/ED visit	0.62	number of opioid dispensations in previous 30 days	0.86
history of alcohol use disorder	0.62	number of BZD dispensations in previous 30 days	0.46
history of fluid and electrolyte disorder	0.32	Doctor risk score*	0.45
history of poisoning	0.31	total OME consumed in previous 30 days	0.43
history of psychoses	0.31	history of poisoning	0.37
number of unique BZD dispensed in previous 30 days	0.26	pharmacy risk score**	0.35
history of depression	0.19	number of unique providers that prescribed an opioid or BZD	0.34
concurrent use of opioid and BZD in previous 30 days	0.19	income	0.34
history of injury	0.17	history of prior hospitalization/ED visit	0.26

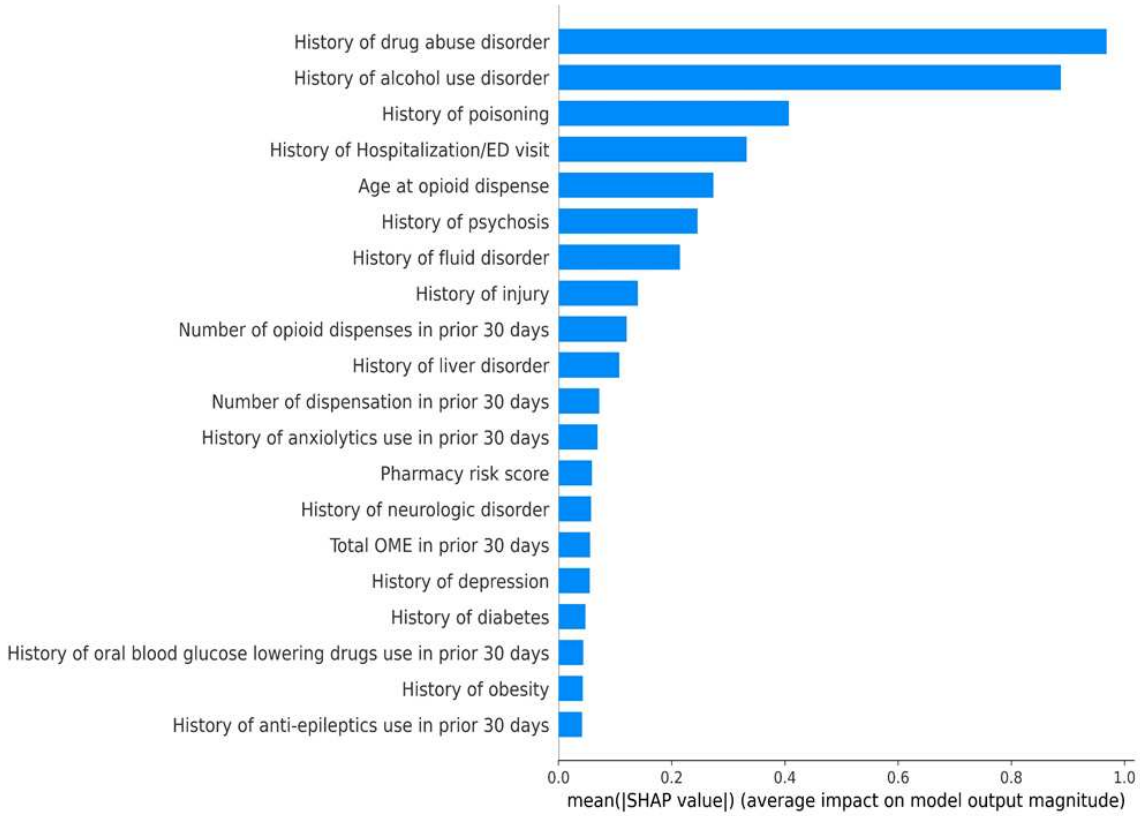
Note: Logistic regression used L1 (lasso) parameter regularization; BZD- benzodiazepine; OME- oral morphine equivalents; ED: emergency department

*derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each physician;

**derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy

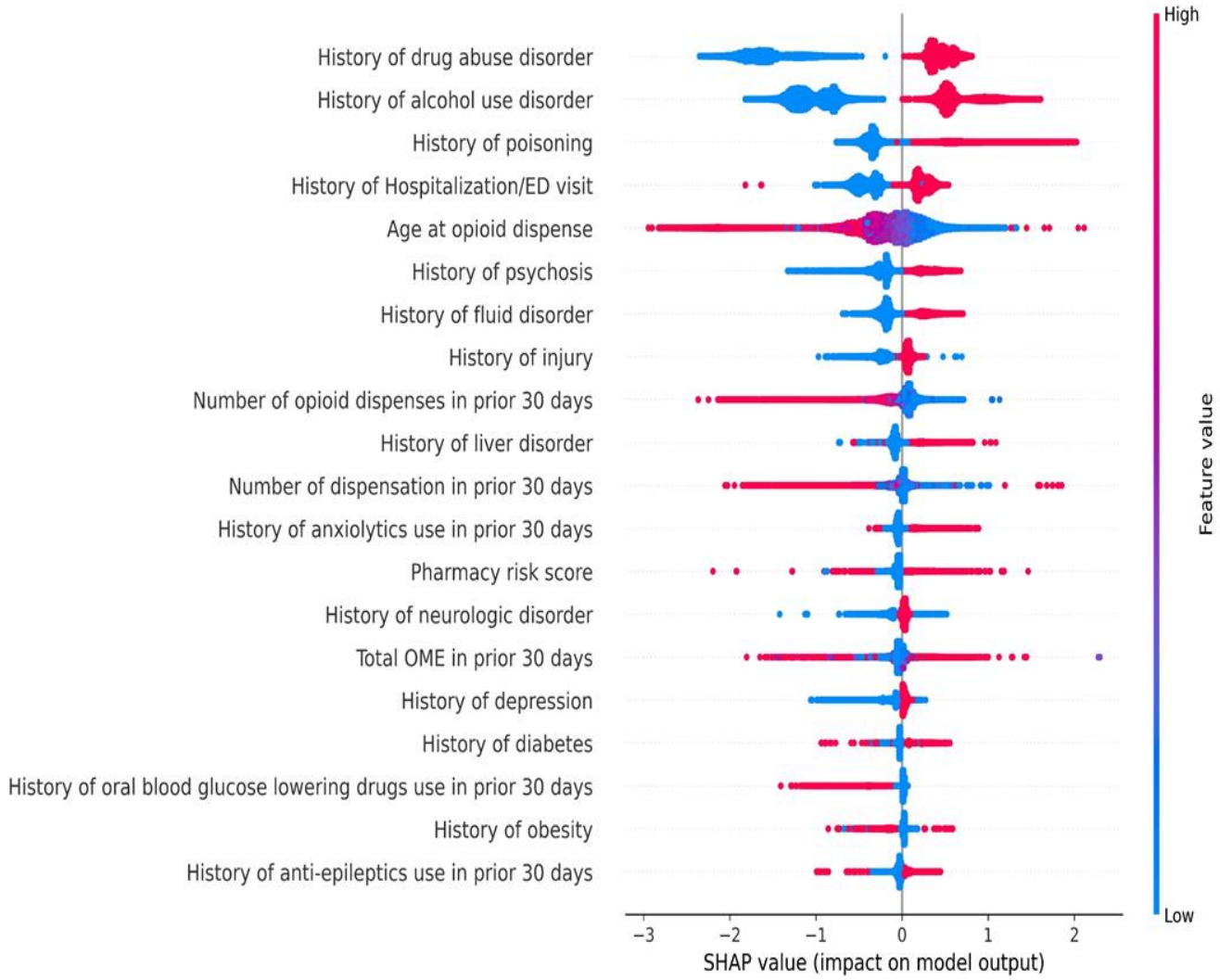
eFigure 3. SHAP values and feature impact of the XGBoost classifier using the 2018 validation set to describe "associations" between features and the outcome. Features with the most impact on the model with drug abuse ranked highest (A); tornado plot illustrating feature impact (B); explaining the prediction of study outcome based on predictor values for 4 patients using SHAP values(C).

(A)



Note: Pharmacy risk score- derived feature using proportion of opioid patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy; training and validating the XGBoost classifier with these features alone resulted in an AUC of 0.877 in the 2018 validation set

(B)

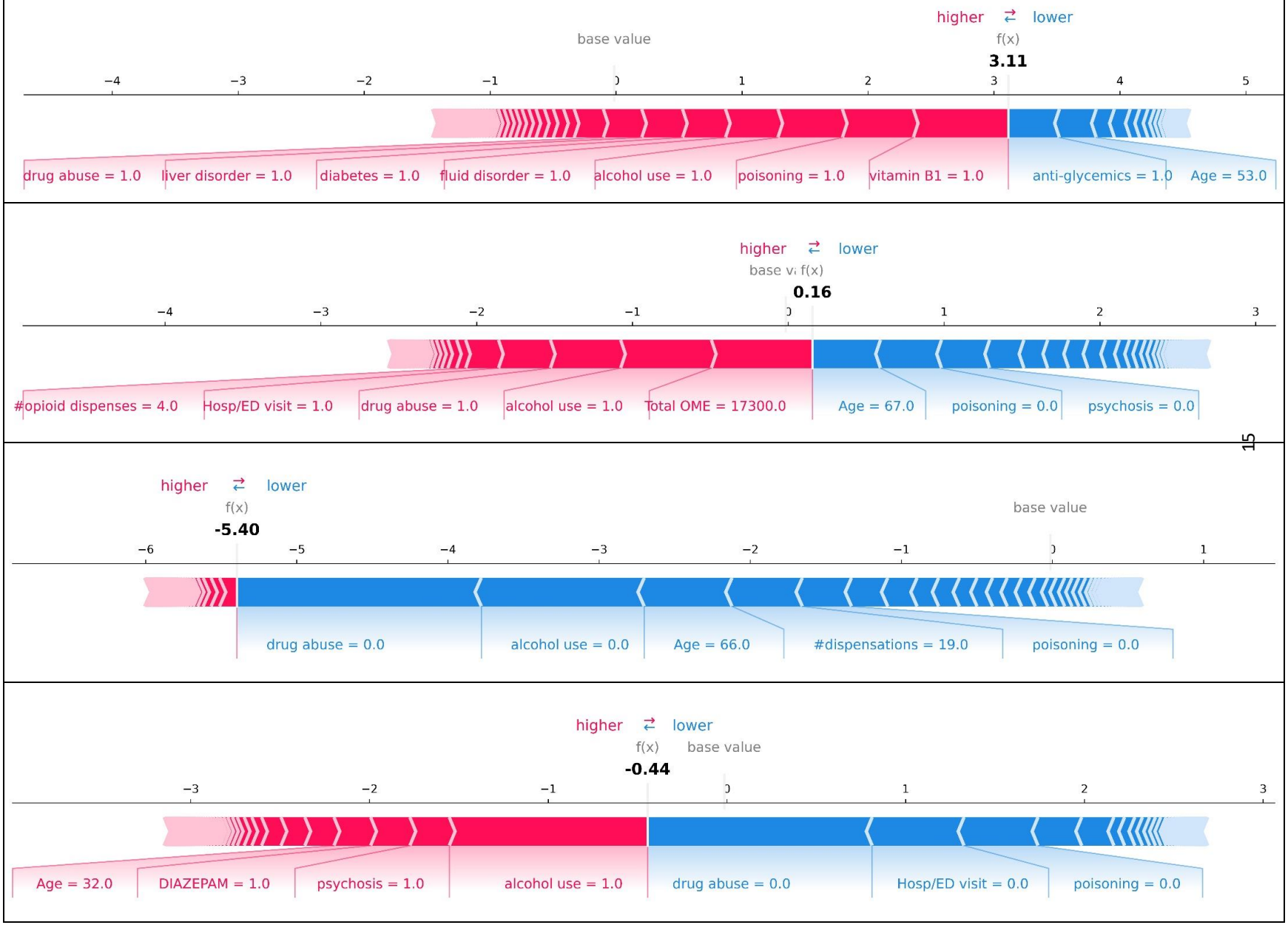


Note: Pharmacy risk score- derived feature using proportion of opioid/benzodiazepine patients that experienced the study outcome in the previous 30 days prior to opioid dispensation for each pharmacy; red indicates higher values of categorical variables and plots to the right of 0.0 indicate the tendency to be associated with the study outcome while blue indicates lower values of categorical variables and plots to the left of 0.0 indicate the tendency to be associated with no outcome

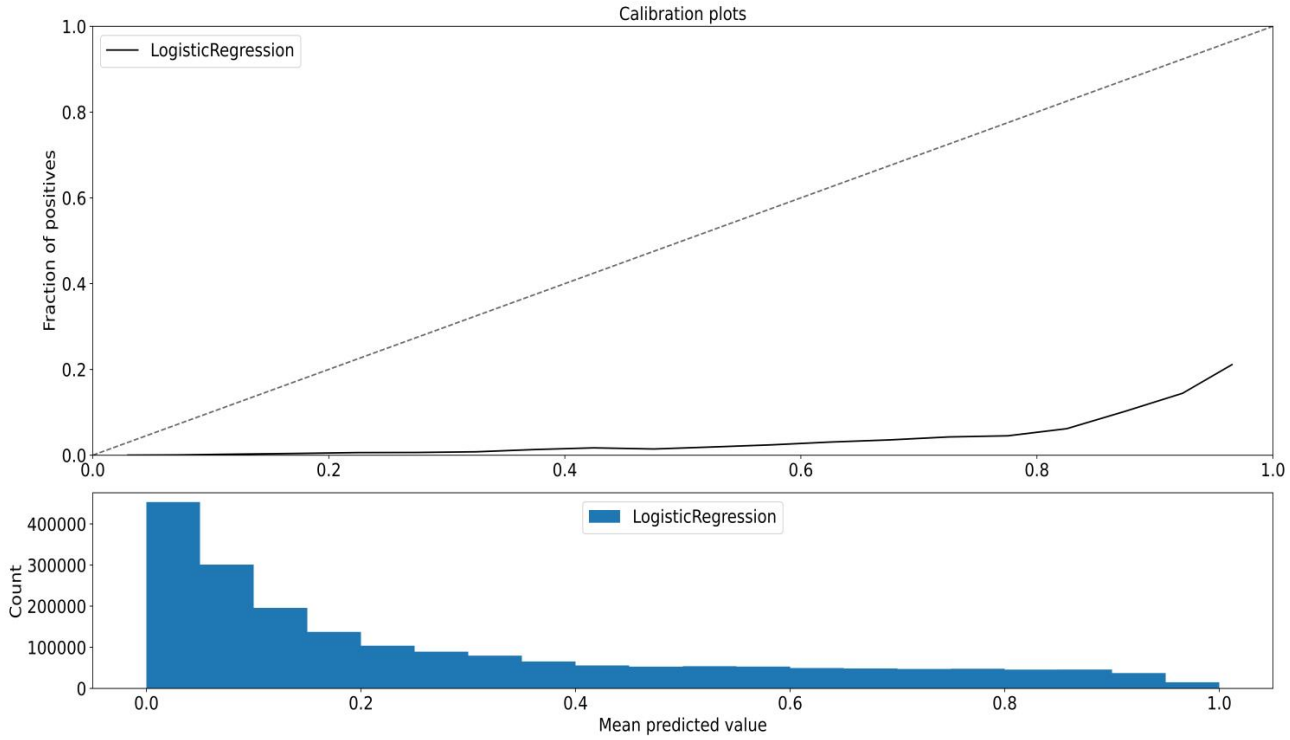
(C)

How to read the figure on the next page: Using hospitalization within 30-days of an opioid dispensation as the outcome of interest, there are 4 scenarios to consider: the XGBoost classifier has low or high confidence in predicting a hospitalization and low or high confidence in predicting **NO** hospitalization. Start at the base SHAP value of near 0.0 ("base value") in which the classifier is not confident in the prediction. SHAP values (in bold) that are above 0.0 indicate a tendency towards a hospitalization while those that are below 0.0 indicate a tendency for **NO** hospitalization. As the SHAP value moves above 0.0, for example 3.11 in the top panel, the classifier's confidence in predicting a hospitalization is higher. As the SHAP value approaches closer to the base value, for example 0.16 in the second panel, the classifier has relatively lower confidence in predicting a hospitalization. When the SHAP value is below 0.0, for example -5.4 in the third panel, the classifier's confidence in predicting **NO** hospitalization is higher and when the SHAP value is closer to 0.0, for example -0.44 in the bottom panel, the classifier has lower confidence in predicting **NO** hospitalization.
The top panel (SHAP value 3.11) depicts an instance predicted to be high risk for our outcome. This individual has a positive history of drug abuse disorder, liver disorder, diabetes, fluid/electrolyte disorder, alcohol use disorder, poisoning and B vitamin use in the prior 30 days. The third panel (SHAP value -5.40) depicts an instance predicted to be low risk (i.e., no hospitalization) and has a negative history for poisoning, drug and alcohol use disorder.
Note- drug abuse: drug abuse disorder; poisoning: history of poisoning; vitamin B1: vitamin B1 in prior 30 days; anti-glycemics: anti-glycemic agents in prior 30 days; age: age at opioid dispensation; # opioid dispenses: number of opioid dispensations in prior 30 days; Hosp/ED visit: history of prior hospitalizations and/or emergency visits in past 6 months; Total OME: total oral morphine equivalents in prior 30 days; DIAZEPAM: history of diazepam use in prior 30 days.

eFigure 3C


higher ⇄ lower
base value
f(x)
3.11
drug abuse = 1.0
liver disorder = 1.0
diabetes = 1.0
fluid disorder = 1.0
alcohol use = 1.0
poisoning = 1.0
vitamin B1 = 1.0
anti-glycemics = 1.0
Age = 53.0
higher ⇄ lower
0.16
#opioid dispenses = 4.0
Hosp/ED visit = 1.0
drug abuse = 1.0
alcohol use = 1.0
Total OME = 17300.0
Age = 67.0
poisoning = 0.0
psychosis = 0.0
higher ⇄ lower
f(x)
-5.40
base value
drug abuse = 0.0
alcohol use = 0.0
Age = 66.0
#dispensations = 19.0
poisoning = 0.0
higher ⇄ lower
f(x)
base value
-0.44
Age = 32.0
DIAZEPAM = 1.0
psychosis = 1.0
alcohol use = 1.0
drug abuse = 0.0
Hosp/ED visit = 0.0
poisoning = 0.0

eFigure 4. Calibration curve plotting observed vs. quantiles of estimated risk for the logistic regression (L1) classifier using the 2018 validation dataset. The majority of counts (dispensations) were predicted to be lower risk.



eReferences

1. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning, vol. 1. Springer series in statistics New York (2001)
2. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
3. Zhu, H. Zou, S. Rosset, T. Hastie, "Multi-class AdaBoost", 2009.
4. Chang CC, Lin CJ. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011 May 6;2(3):1-27.
5. Scikit-learn: Machine Learning in Python, Pedregosa et al., JMLR 12, pp. 2825-2830, 2011.
6. Nair V, Hinton GE. Rectified linear units improve restricted boltzmann machines. InProceedings of the 27th international conference on machine learning (ICML-10) 2010 (pp. 807-814).
7. Chen T, Guestrin C. Xgboost: A scalable tree boosting system. InProceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining 2016 Aug 13 (pp. 785-794).