THE LANCET
Digital Health

Supplementary appendix

This appendix formed part of the original submission and has been peer reviewed. We post it as supplied by the authors.

Supplement to: Lo-Ciganic W-H, Donohue JM, Yang Q, et al. Developing and validating a machine-learning algorithm to predict opioid overdose in Medicaid beneficiaries in two US states: a prognostic modelling study. *Lancet Digit Health* 2022; **4:** e455–65.

Appendix Table of Contents

Appendix Methods

Developing prediction algorithm using 2013-2016 Pennsylvania Medicaid data

In this study, our primary goal was prediction, and our secondary goal was risk stratification (i.e., to identify subgroups of patients at similar risk of the outcome). First, we randomly and equally divided the 2013-2016 Pennsylvania Medicaid beneficiaries into training, testing, and validation samples based on the beneficiaries' characteristics and opioid overdose distribution. We created a series of candidate predictors (n=284) identified from prior literature(1-24) and our previous work (**Appendix p5**)(25) that were measured at baseline (during the 3-month period before the first opioid fill) and in 3-month windows after initiating prescription opioids. **Appendix p5** lists each of the candidate predictors related to health status (e.g., number of ED visits, comorbidities), patterns of opioid (e.g., total morphine milligram equivalent) and other relevant medication (e.g., benzodiazepines) use, regional-level factors linked from publicly-available sources (e.g., area deprivation index), and provider-level variables (e.g., specialty).(26) We used the sliding-window and multi-instance approach that was conceptually similar to discrete-time survival analysis methods in which covariates are processed in sequential chunks.(27, 28) This approach better simulates continuous population screening in practical applications compared to time-series analysis. We simulated a system in which the entire cohort was screened every 3 months to accurately capture all instances of overdoses during the target prediction window. We aimed to answer the question: "Will the patient have an overdose event at any time point during the target subsequent 3-month window?". Beneficiaries remained in the cohort once eligible, regardless of whether they continued to receive opioids or had an overdose, until they died or disenrolled from Medicaid programs. Machine learning can handle highly correlated data with repeated opioid episodes or outcome events per patient. We developed and tested prediction algorithms for the risk of opioid overdose using gradient boosting machine (GBM). We fitted the trained algorithms based on the training sample, refined the algorithm using the testing sample, and then applied the final algorithm to the validation sample to evaluate prediction performance.

Our model reporting complies with the Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) and the Standards for Reporting of Diagnostic Accuracy (STARD) reporting guidelines.(27, 28) According to the TRIPOD guideline, multivariable prediction models fall into 2 broad categories (1) diagnostic and (2) prognostic prediction models. A diagnostic modeling study includes multiple predictor candidates to estimate the probability that a certain condition or disease is present (or absent) at the moment of prediction (i.e., cross-sectional design). A prognostic modeling study includes multiple predictor candidates to estimate the probability of a particular outcome occurring in a certain period in the future (e.g., overdose in the subsequent 3 months in our study). Our study was a prognostic modeling study (with a retrospective longitudinal design). We calculated the C-statistic (or the area under the receiver operating curve [ROC]) from the validation sample to assess discrimination (i.e., the extent to which patients predicted as high-risk exhibit higher overdose rates compared to those predicted as low-risk). For each probability cutoff point, opioid overdose was predicted for the visits with calculated probabilities above the cutoff point, whereas non-overdose was predicted for the visits with probabilities below the cutoff point. Based on their true and predicted opioid overdose status, the patients' 90-day visits can be assigned to one of the four groups (i.e., true positive [TP], false positive [FP], true negative [TN], false negative [FN]) as shown in the classification matrix (**Appendix p12-13**). Given that opioid overdose events are rare outcomes and C-statistics do not incorporate information about the prevalence of the outcome, we further reported other more appropriate metrics, including sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), positive likelihood ratio (PLR), negative likelihood ratio (NLR), number needed to evaluate (NNE) to identify one opioid overdose, and estimated rate of alerts to assess pre-implementation evaluation of our prediction algorithms (**Appendix p12-13**).(29) The optimal algorithm for a screening test depends on the pre-test probability of the outcome, the values of TPs and

TNs, and the costs of FP and FN. Since these factors vary from setting to setting (and some of them are subjective choices), no single cutoff point is suitable for every purpose. To compare performance across methods, we presented and assessed these prediction metrics (e.g., NNE) at the balanced threshold of the predicted probability that balances sensitivity and specificity as identified by the Youden index,(30) as well as at multiple levels of sensitivity and specificity (e.g., 90%-100%) to allow risk-benefit evaluations of interventions triggered by positive tests using different thresholds to define high risk.

Second, based on the individual's predicted probability of an opioid overdose event, we classified beneficiaries in the validation sample into decile risk subgroups using the risk score thresholds derived from the 2013-2016 Pennsylvania training algorithm, with the highest decile further split into three additional strata based on the top 1st, 2nd to 5th, and 6th to 10th percentiles to allow closer examination of patients at highest risk of experiencing opioid overdose. We evaluated calibration plots (the extent to which the predicted opioid overdose risk agreed with the observed risk) by risk subgroup. We briefly summarized the GBM approach in the sections below (see more details in our previously published work).(25) **Gradient Boosting Machine (GBM; Stochastic gradient boosting or TreeNet in Salford SPM)(31, 32)**

GBM, a tree-structured ensemble approach, consists of a series of trees grown in a sequential order of successive trees to minimize the residual error. We used the Salford's TreeNet function to supply an initial value specific to the chosen loss function (i.e. logistic binary) for each record in the training sample. The Salford's TreeNet function is similar to the XGBoost in Python, which can handle a large number of features. TreeNet can handle missing values automatically. No additional feature selection process was used prior to the GBM modeling. We used both cross entropy (i.e., negative average log likelihood) and the area under the receiver operating characteristic curve (AUCROC) methods as the tuning criterion to determine the optimal number of trees optimal for logistic models. The Neg. AvgLL approach is similar to AUCROC, but emphasizes the probability interpretation of the model predictions. The AUCROC is a measure the overall model performance tied closely to the ability of the model to correctly rank records from most likely to least likely to be "1" or "0". Both approaches yielded similar optimal numbers of trees, and we reported the results from the Neg. AvgLL approach because relying on the C-statistic (i.e., AUCROC) can be misleading for rare outcomes. Second, TreeNet sampled 25% of the records in the training sample randomly (4-fold cross-validation) and then computed the generalized residual for the records in the sample. The first tree is fitted to the data and begins with a very small tree as the initial model. TreeNet used the sampled records to fit a classification tree with a maximum of 8 terminal nodes to the generalized residuals. Third, TreeNet used the classification tree derived from the sampled records to update the TreeNet model based on the loss function and shrank the updated tree by the best learning rate (or the shrinkage rate) at 0.1 for overfitting protection. TreeNet repeated the steps previously described 50 to 300 times (i.e., the best number of trees to build = 200). Other parameters used for tuning included maximum depth of the tree (3-8), feature resample rate (i.e. columns) (0.7-0.9), data resample rate (i.e. rows) (0.7-0.9), L1 regularization weight (0.01-1), L2 regularization weight (0.01-1), minimum child weight (to prevent further partition, i.e., overfitting) (0.01-1), minimum loss reduction (required to make further partition) (1-50), and step size shrinkage [0.1-0.5]. Finally, we tested and validated the algorithms in the testing and validation samples. For studies like ours with a highly imbalanced classification, calibration of the probability is needed to avoid being over-confident about the prediction performance. To obtain the true event probability, we fit a logistic regression model ($Y=\beta 0+\beta 1\times$predicted ½ log-odds of GBM's predicted score) to transform the calibrated scores.

eTable 1. Diagnosis codes for identifying opioid overdose

Conditions	ICD-9 codes	ICD-10 codes
Opioid overdose	965.00, 965.01, 965.02, 965.09, E850.0, E850.1, E850.2, E935.0, E935.1, E935.2	T40.0X1A, T40.0X2A, T40.0X3A, T40.0X4A, T40.1X1A, T40.1X2A, T40.1X3A, T40.1X4A, T40.2X1A. T40.2X2A, T40.2X3A, T40.2X4A, T40.3X1A, T40.3X2A, T40.3X3A, T40.3X4A, T40.4X1A, T40.4X2A, T40.4X3A, T40.4X4A, T40.601A, T40.602A, T40.603A, T40.604A, T40.691A, T40.692A, T40.693A, T40.694A

eTable 2. Other diagnosis codes used to identify the likelihood of opioid overdose[a]

ICD type	ICD code	ICD codes description
Other drug/substance-related overdose or substance use disorders		
ICD-9	965*	Poisoning by analgesics antipyretics and anti-rheumatics
ICD-9	966	Poisoning by anticonvulsants and anti-parkinsonism drugs
ICD-9	967	Poisoning by sedatives and hypnotics
ICD-9	968	Poisoning by other central nervous system depressants and anesthetics
ICD-9	969	Poisoning by psychotropic agents
ICD-9	970	Poisoning by central nervous system stimulants
ICD-9	971	Poisoning by drugs primarily affecting the autonomic nervous system
ICD-9	972	Poisoning by agents primarily affecting the cardiovascular system
ICD-9	973	Poisoning by agents primarily affecting the gastrointestinal system
ICD-9	975	Poisoning by agents primarily acting on the smooth and skeletal muscles and respiratory system
ICD-9	977	Poisoning by other and unspecified drugs and medicinal substances
ICD-9	980	Toxic effect of alcohol
ICD-9	989	Toxic effect of other substances chiefly nonmedicinal as to source
ICD-9	303	Alcohol dependence syndrome
ICD-9	304	Drug dependence
ICD-9	305	Nondependent abuse of drugs
ICD-10	F10	Alcohol related disorders
ICD-10	F11	Opioid related disorders
ICD-10	F12	Cannabis related disorders
ICD-10	F13	Sedative, hypnotic, or anxiolytic related disorders
ICD-10	F14	Cocaine related disorders
ICD-10	F15	Other stimulant related disorders
ICD-10	F16	Hallucinogen related disorders
ICD-10	F17	Nicotine dependence
ICD-10	F18	Inhalant related disorders
ICD-10	F19	Other psychoactive substance related disorders
ICD-10	T39	Poisoning by, adverse effect of and underdosing of nonopioid analgesics, antipyretics and antirheumatics
ICD-10	T40	Poisoning by, adverse effect of and underdosing of narcotics and psychodysleptics [hallucinogens]
ICD-10	T41	Poisoning by, adverse effect of and underdosing of anesthetics and therapeutic gases
ICD-10	T42	Poisoning by, adverse effect of and underdosing of antiepileptic, sedative- hypnotic and antiparkinsonism drugs
ICD-10	T43	Poisoning by, adverse effect of and underdosing of psychotropic drugs, not elsewhere classified
ICD-10	T48	Poisoning by, adverse effect of and underdosing of agents primarily acting on smooth and skeletal muscles and the respiratory system
ICD-10	T51	Toxic effect of alcohol
ICD-10	T65	Toxic effect of other and unspecified substances

* Excluding codes for opioid and heroin overdose.
[a:] Based on Dunn KM et al. (2010)(7) but excluding E950-959 (suicide and self-inflicted injury codes).

Lo-Ciganic et al. Using machine learning to predict opioid overdose in Medicaid

eTable 3. Summary of predictor candidates (n=284) measured in 3-month windows for predicting subsequent opioid overdose[a]

Patterns of prescription opioid use[b]	Patterns of non-opioid prescription use	Beneficiaries sociodemographics	Health status factors	Opioid prescriber-level variables (PA Medicaid only)[d]	Regional-level factors[e]
• Average opioid daily dose in MME[c] • Cumulative MME • Cumulative duration for any opioids, SAO, and LAO • Duration of longest continuous use for any opioids, SAO, and LAO • No. fills of any opioids, SAO, and LAO • No. standardized 30-day prescriptions for any opioids, SAO, and LAO • Cumulative duration of 30-day use of any opioids, SAO, and LAO • No. fills by opioid ingredient and type (e.g., any fentanyl, SAO-type fentanyl, LAO-type fentanyl) • Type of opioids by Schedule and SAO/LAO (e.g., SAO, Schedule I only) • No. unique opioid prescribers • No. unique pharmacies • No. early refills for opioids • Cumulative overlapping days of early refills	• No. BZD fills • No. muscle relaxants fills • Cumulative overlapping days of concurrent opioid and BZD use • Cumulative overlapping days of concurrent opioid and muscle relaxants use • Cumulative overlapping days of concurrent opioid, BZD and muscle relaxants use • Cumulative duration of naltrexone • No. gabapentinoid fills • Cumulative duration of gabapentinoid use • No. antidepressants fills • Cumulative duration of antidepressant use • No. average monthly non-opioid prescriptions • No. naltrexone fills • Received methadone opioid agonist therapy[f] • Received buprenorphine for OUD[f] • Cumulative duration of buprenorphine for OUD[f]	• Age • Sex • Race (White, Black, Other/Unknown) • Ethnicity (Hispanic, non-Hispanic, and Other/Unknown)[g] • County of residence • Zip code of residence • Type of resided county (metro vs. non-metro) • Type of Medicaid eligibility • Duration of Medicaid enrollment	• No. outpatient visits • No. ED visits • No. inpatient visits • History of prescription opioid overdoses • History of heroin overdose • History of naloxone administration • Non-opioid drug use disorders • Alcohol use disorders • History of urine drug tests • History of SUD counseling • OUD • Adjustment disorders • Personality disorders • Psychoses • Delusional disorders • Schizophrenia • Mood disorders • Anxiety disorders • Alcohol-induced mental disorders • Drug-induced mental or sleep disorders • Other mental health disorders • Osteoarthritis • Rheumatoid arthritis • Back pain • Neck pain • Headache or migraine • Temporomandibular disorder pain • Abdominal pain or hernia • Chest pain • Kidney or gall bladder stones • Menstrual or genital reproductive pain • Fractures, concussion, injuries • Fibromyalgia • Internal orthopedic device implant/graft • Other pain conditions • Surgical procedures (e.g., ischemic heart diseases) • Diseases of musculoskeletal system and connective tissues • Neuropathies (excluding alcoholic, drug, and optic related) • Ischemic heart disease • HIV/AIDS • Elixhauser index and individual categories	• Prescriber's sex • Prescriber's specialties • Average monthly opioid prescribing volume • Average monthly opioid prescribing dose in MME • Average monthly No. of patients receiving opioids	• AHRF total health facilities variables • AHRF health professions variables • AHRF resource scarcity variables • AHRF health training programs variables • AHRF hospital expenditures, Medicare costs, VA expenditures • AHRF inpatient days/discharges variables • AHRF other health services utilization variables • AHRF census-based variables (e.g., medium household income, employment) • AHRF health insurance status variables • AHRF housing statistics • County health rankings and roadmaps • Area deprivation index County-health ranking variables

Abbreviations: AHRF: Area Health Resources Files; **BZD:** benzodiazepines; **DUI:** driving under the influence**; HIV/AIDS:** human immunodeficiency virus/acquired immunodeficiency syndrome; **LAO:** long-acting opioids**; MME:** morphine milligram equivalent; **No**: Number of; **SAO:** short-acting opioids**; SUD:** substance use disorders;

[a:] Details for the operational definitions for each variable and corresponding diagnosis and procedure codes and National Drug Codes can be provided by request to the corresponding author.

[b:] We used an "as-prescribed" approach that assumes patients taking all prescribed opioids on the schedule recommended by their clinicians.(10) Patients who received refills for the same drug at the same dose and schedule while still having opioid prescriptions within three days from a prior fill were assumed to have taken the medication from the prior fill before taking medication from the second fill.(33)

[c:] We calculated morphine milligram equivalent (MME) for each opioid prescription, defined by the quantity dispensed multiplied by the strength in milligrams, multiplied by a conversion factor.(34) For each person, the average daily MME during the 90-day window was calculated by summing MMEs across all opioids and dividing by the number of days supplied.

[d:] Prescribers were identified by their National Provider Identifiers. Primary opioid prescribers were defined as the prescribers who dominantly prescribed the most opioid prescriptions. If patients only had 2 opioid prescriptions, then the first prescriber was considered as the primary prescriber.

[e:] AHRF variables (https://data.hrsa.gov/topics/health-workforce/ahrf), area deprivation index (https://www.hipxchange.org/ADI), and county-health ranking variables (http://www.countyhealthrankings.org/explore-health-rankings/use-data) are publicly available and downloadable

[g]: Arizona Medicaid data did not have a separate ethnicity variable from race, we create the ethnicity variable and classified beneficiaries as Hispanic when it was indicated in the race category or death certificates as Hispanic; otherwise, we classified the remaining Arizona Medicaid beneficiaries as non-Hispanic.

eTable 4. Prediction performance measures for predicting opioid overdose (fatal/nonfatal) varying sensitivity and specificity using gradient boosting machine (GBM): 2013-2016 and 2017-2018 Pennsylvania (PA) Medicaid data and Arizona (AZ) Medicaid claims

Methods	Score threshold (range 0-100)[a]	Predicted overdose (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1 (%)	F2 (%)	PLR	NNE
2013-2016 PA data										
Sensitivity										
100%	7.93	99.70	100.00	0.30	0.19	100.00	0.0039	0.0096	1.00	516
99%	15.82	88.56	99.02	11.46	0.22	99.98	0.0043	0.0107	1.12	463
98%	18.91	78.83	98.02	21.20	0.24	99.98	0.0048	0.0119	1.24	417
97%	21.14	71.48	97.01	28.57	0.26	99.98	0.0052	0.0130	1.36	382
96%	22.91	66.08	96.00	33.98	0.28	99.98	0.0056	0.0139	1.45	357
95%	24.73	60.94	95.03	39.12	0.30	99.98	0.0060	0.0149	1.56	332
94%	26.17	56.89	94.02	43.18	0.32	99.97	0.0064	0.0157	1.65	313
93%	27.55	53.27	93.01	46.80	0.34	99.97	0.0067	0.0166	1.75	297
92%	28.43	51.09	92.00	48.99	0.35	99.97	0.0069	0.0171	1.80	288
91%	29.58	48.27	91.03	51.81	0.36	99.97	0.0073	0.0179	1.89	275
90%	30.52	46.12	90.02	53.96	0.38	99.96	0.0075	0.0185	1.96	265
85%	35.91	35.48	85.02	64.61	0.46	99.96	0.0092	0.0226	2.40	216
80%	41.12	27.93	80.01	7.17	0.55	99.95	0.0110	0.0269	2.88	181
75%	46.97	21.49	75.01	78.62	0.67	99.94	0.0134	0.0325	3.51	148
Balanced threshold[b]	46.79	21.65	75.27	78.46	0.67	99.94	0.0133	0.0324	3.49	149
Specificity										
90%	65.20	10.09	59.16	90.00	1.13	99.91	0.0222	0.0526	5.92	88
91%	67.96	9.09	56.47	91.00	1.20	99.91	0.0235	0.0553	6.27	83
92%	71.45	8.09	52.92	92.00	1.26	99.90	0.0247	0.0577	6.61	79
93%	75.96	7.08	49.36	93.00	1.35	99.89	0.0262	0.0607	7.05	74
94%	83.05	6.07	44.71	94.00	1.42	99.89	0.0275	0.0630	7.45	70
95%	96.70	5.07	38.97	95.00	1.49	99.88	0.0286	0.0644	7.79	67
96%	97.45	4.06	34.05	96.00	1.62	99.87	0.0309	0.0681	8.51	62
97%	97.84	3.05	28.01	97.00	1.77	99.86	0.0334	0.0708	9.34	56
98%	98.13	2.04	20.40	98.00	1.94	99.84	0.0353	0.0701	10.20	52
99%	98.40	1.02	11.96	99.00	2.26	99.83	0.0381	0.0644	11.97	44
100%	99.48	0.00	0.00	100.00	0.00	99.81	N/A	N/A	0.00	inf
Maximized PPV	99.17	0.00	0.06	100.00	5.26	99.81	0.0012	0.0007	28.72	19
2017-2018 PA data										
Sensitivity										
100%	8.22	99.76	100.00	0.24	0.17	100.00	0.0033	0.0083	1.00	602
99%	15.98	90.64	99.01	9.37	0.18	99.98	0.0036	0.0090	1.09	552
98%	18.78	83.14	98.02	16.89	0.20	99.98	0.0039	0.0097	1.18	512
97%	20.88	76.89	97.03	23.14	0.21	99.98	0.0042	0.0104	1.26	478
96%	23.08	70.57	96.01	29.47	0.23	99.98	0.0045	0.0112	1.36	443
95%	24.45	66.71	95.02	33.34	0.24	99.98	0.0047	0.0117	1.43	423
94%	25.88	62.65	94.03	37.40	0.25	99.97	0.0050	0.0123	1.50	402
93%	27.11	59.23	93.01	40.82	0.26	99.97	0.0052	0.0129	1.57	384
92%	28.22	56.22	92.02	43.84	0.27	99.97	0.0054	0.0134	1.64	369
91%	29.36	53.30	91.03	46.77	0.28	99.97	0.0056	0.0140	1.71	353
90%	30.34	50.88	90.01	49.18	0.29	99.97	0.0058	0.0145	1.77	341
85%	35.56	39.74	85.03	60.34	0.35	99.96	0.0071	0.0174	2.14	282
80%	41.06	30.39	80.01	69.70	0.44	99.95	0.0087	0.0214	2.64	229
75%	45.35	24.54	75.03	75.54	0.51	99.95	0.0101	0.0247	3.07	197
Balanced threshold[a]	49.30	20.19	71.37	79.89	0.59	99.94	0.0116	0.0284	3.55	171
Specificity										
90%	63.90	10.08	56.50	90.00	0.93	99.92	0.0183	0.0436	5.65	108
91%	66.54	9.07	53.86	91.01	0.98	99.92	0.0193	0.0459	5.99	102
92%	69.70	8.07	51.06	92.00	1.05	99.91	0.0206	0.0485	6.38	95
93%	74.09	7.07	48.05	93.00	1.13	99.91	0.0220	0.0515	6.87	89
94%	81.35	6.06	43.83	94.01	1.20	99.90	0.0234	0.0541	7.31	83
95%	95.55	5.06	39.12	95.00	1.28	99.89	0.0248	0.0567	7.83	78
96%	96.49	4.05	35.72	96.00	1.46	99.89	0.0281	0.0628	8.93	68
97%	97.02	3.05	30.67	97.00	1.67	99.88	0.0317	0.0686	10.23	60
98%	97.43	2.04	24.70	98.00	2.01	99.87	0.0372	0.0758	12.35	50
99%	97.86	1.02	14.51	99.00	2.36	99.86	0.0405	0.0714	14.53	42
100%	99.27	0.00	0.00	100.00	0.00	99.83	N/A	N/A	0.00	inf

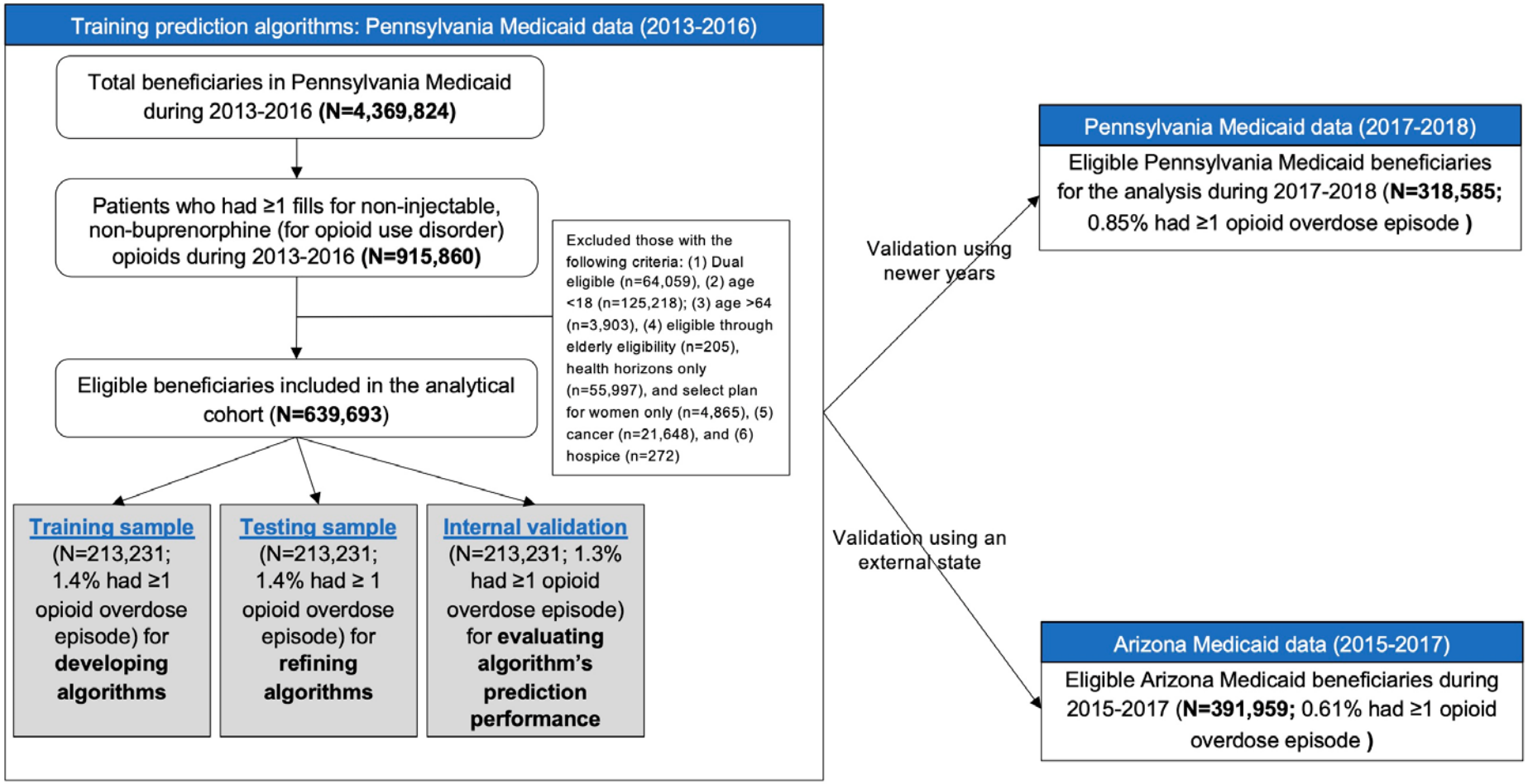
Maximized PPV	99.00	0.00	0.16	100.00	6.10	99.83	0.0032	0.0020	39.10	16
2015-2017 AZ data										
Sensitivity										
100%	6.87	99.65	100.00	0.35	0.09	100.00	0.0017	0.0043	1.00	1170
99%	10.89	93.13	99.03	6.88	0.09	99.99	0.0018	0.0045	1.06	1105
98%	12.24	88.11	98.02	11.90	0.09	99.99	0.0019	0.0047	1.11	1056
97%	13.16	83.92	97.01	16.09	0.10	99.98	0.0020	0.0049	1.16	1016
96%	13.82	80.23	96.04	19.79	0.10	99.98	0.0020	0.0051	1.20	981
95%	14.94	74.24	95.03	25.77	0.11	99.98	0.0022	0.0054	1.28	918
94%	15.59	70.51	94.01	29.51	0.11	99.98	0.0023	0.0056	1.33	881
93%	16.43	65.77	93.00	34.25	0.12	99.98	0.0024	0.0060	1.41	831
92%	17.06	62.57	92.03	37.45	0.13	99.98	0.0025	0.0062	1.47	799
91%	17.61	59.68	91.02	40.35	0.13	99.98	0.0026	0.0065	1.53	770
90%	18.16	57.03	90.01	42.99	0.13	99.98	0.0027	0.0067	1.58	744
85%	20.76	45.27	85.04	54.76	0.16	99.98	0.0032	0.0079	1.88	625
80%	23.40	35.72	80.02	64.32	0.19	99.97	0.0038	0.0094	2.24	524
75%	27.22	26.46	75.05	73.58	0.24	99.97	0.0048	0.0119	2.84	414
Balanced threshold[a]	34.78	16.06	67.17	83.99	0.36	99.97	0.0071	0.0174	4.19	281
Specificity										
90%	43.06	10.04	58.56	90.00	0.50	99.96	0.0098	0.0240	5.86	201
91%	45.17	9.04	56.58	91.00	0.53	99.96	0.0106	0.0257	6.29	188
92%	47.65	8.04	54.19	92.00	0.57	99.96	0.0114	0.0275	6.78	174
93%	50.63	7.03	51.52	93.01	0.62	99.96	0.0123	0.0297	7.37	160
94%	54.22	6.04	48.39	94.00	0.68	99.95	0.0135	0.0323	8.07	146
95%	58.98	5.02	44.84	95.01	0.76	99.95	0.0149	0.0356	8.99	132
96%	65.86	4.02	39.46	96.01	0.83	99.95	0.0163	0.0385	9.88	120
97%	93.70	3.02	29.70	97.00	0.84	99.94	0.0163	0.0376	9.90	120
98%	96.03	2.02	23.16	98.00	0.98	99.93	0.0188	0.0418	11.58	102
99%	96.92	1.01	13.77	99.00	1.16	99.93	0.0214	0.0434	13.77	86
100%	99.05	0.00	0.05	100.00	100.00	99.91	0.0009	0.0006	inf	1
Maximized PPV	99.05	0.00	0.05	100.00	100.00	99.91	0.0009	0.0006	inf	1

Abbreviations: **AZ:** Arizona; **GBM**: gradient boosting machine; **INF**: infinity; **N/A**: not able to calculated; **NNE**: number needed to evaluate; **NPV**: negative predictive values; **PA:** Pennsylvania; **PLR**: positive likelihood ratio; **PPV**: positive predictive values; **RF**: random forest.

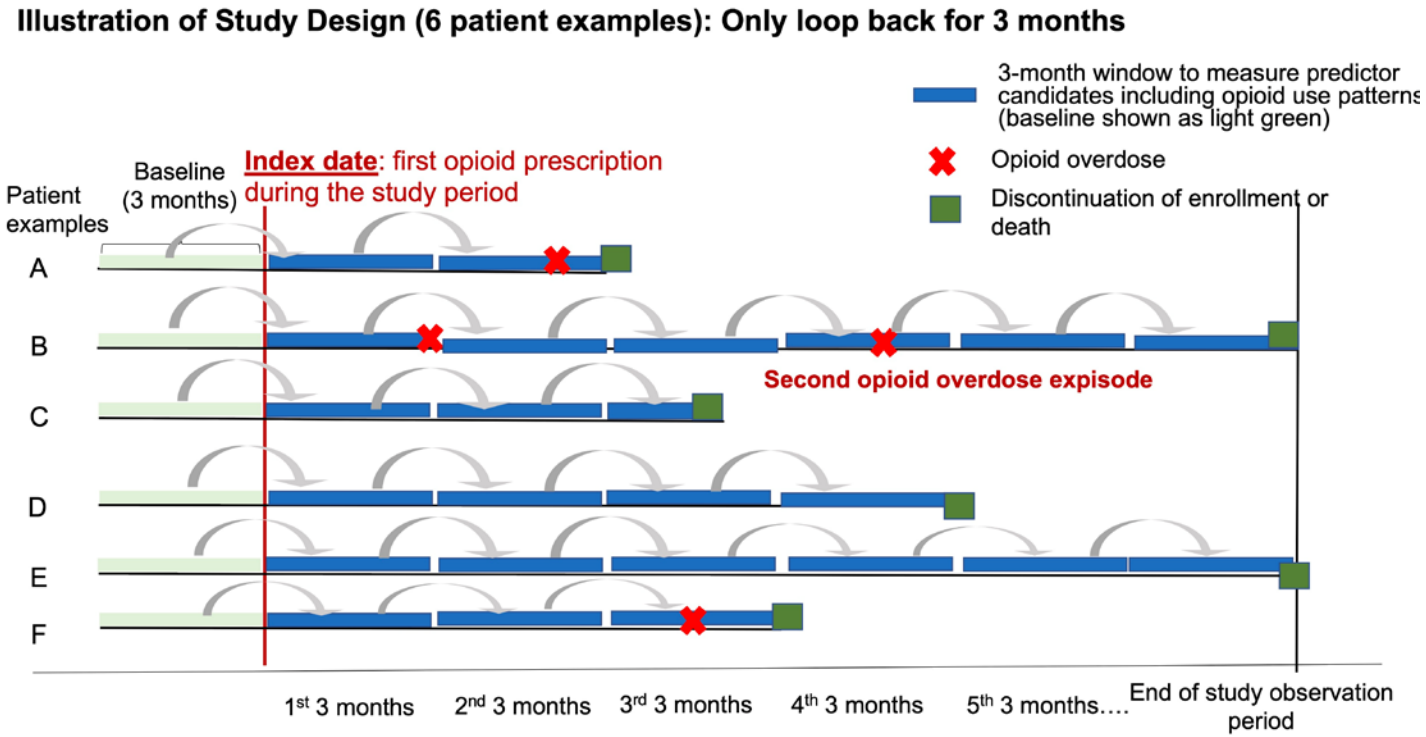
[a]: Scores were calculated by predicted probability multiplied by 100. Score threshold refers to the score used to classify or predict individuals with opioid overdose (i.e., ≥ the threshold) vs. non-overdose (i.e., <threshold)

[b]: Balanced threshold was calculated by the Youden Index to achieve balanced sensitivity and specificity.

eFigure 1. Sample size flow chart of the study cohorts

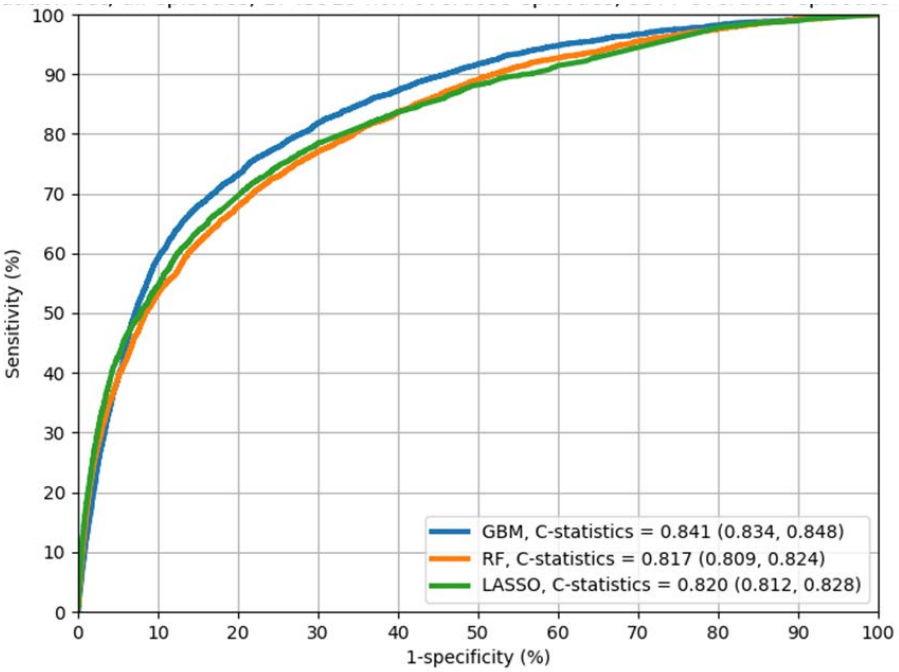


eFigure 2. Study design diagram



Each patient had at least one pair eligible data (i.e., predictors measured in the 3 months with outcomes measured in the subsequent 3 months) during the study period. An index date was defined as the first opioid fill during our study period. We followed patients starting every 3 months after the index date until they were censored because of death or disenrollment. We measured predictor candidates and opioid overdose episodes for the 3-month periods. This sliding-window and multi-instance approach simulates continuous population screening in a practical application (i.e., simulating a system in which the entire cohort was screened every 3 months, and the system's task was to accurately capture all instances of overdoses at any time in the target prediction window).

eFigure 3. C-statistics for predicting opioid overdose using gradient boosting machine (GBM), random forests, and least absolute shrinkage and selection operator (LASSO): 2013-2016 internal validation Pennsylvania Medicaid episode-level data



Abbreviations: AUC: area under the curves; **AZ:** Arizona; **PA:** Pennsylvania; **ROC:** receiver operating characteristic

eFigure 4. Classification matrix and definition of prediction performance metrics

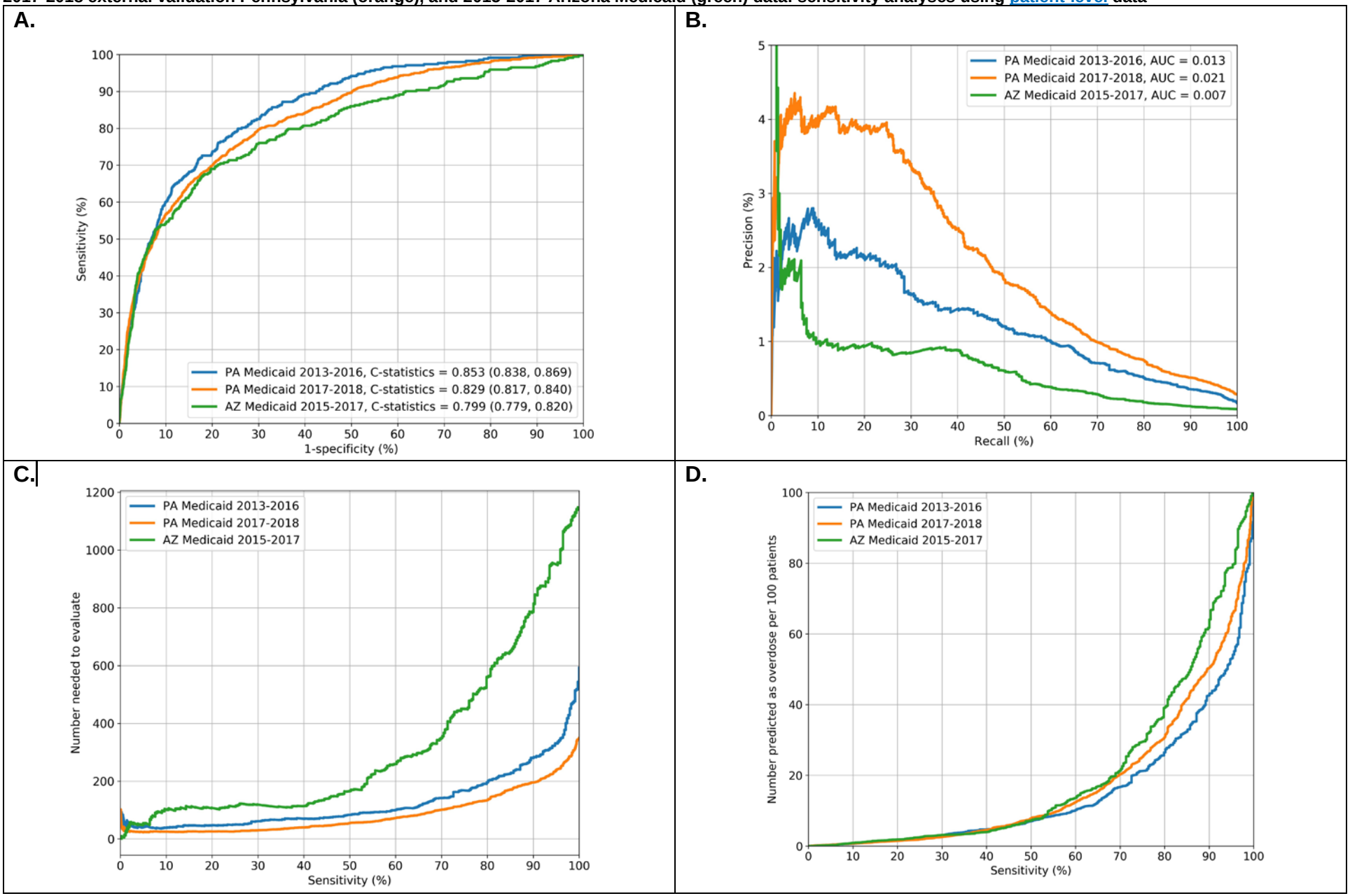
Classification matrix	Predicted category	
Actual Category	Opioid overdose	Non-Opioid overdose
Opioid overdose	True positive (TP)	False negative (FN)
Non-opioid overdose	False positive (FP)	True Negative (TN)

- $\text{Sensitivity}(S_e)\text{ or recall (Rc)} = \frac{TP}{TP+FN}$
- $\text{Specificity}(S_p) = \frac{TN}{FP+TN}$
- $\text{Positive predictive value (PPV) or precision (Pr)} = \frac{TP}{TP+FP} = \frac{\text{sensitivity}\times\text{prevalence}}{\text{sensitivity}\times\text{prevalence}+(1-\text{specificity})\times(1-\text{prevalence})}$
- $\text{Negative predictive value (NPV)} = \frac{TN}{FN+TN}$
- $\text{Overall misclassification rate} = \frac{FP+FN}{TP+FN+FP+TN}$
- $\text{Positive likelihood ratio (PLR)} = \frac{\text{sensitivity}}{1-\text{specificity}}$
- $\text{Negative likelihood ratio (NLR)} = \frac{\text{specificity}}{1-\text{sensitivity}}$
- $\text{F1 score} = 2\frac{Pr\times Rc}{Pr+Rc}$
- $\text{F2 score} = 5\frac{Pr\times Rc}{4\times Pr+Rc}$

Prediction metrics	Definition
Sensitivity (Se) or recall (Rc)	The proportion of correctly predicted positive individuals with opioid overdose (i.e., predicted overdose) divided by all individuals with actual overdose.
Specificity (Sp)	The proportion of correctly predicted negative individuals (i.e., predicted non-overdose) divided by all observations with actual non-overdose.
Positive predictive value (PPV) or precision (Pr)	The proportion of actual opioid overdose cases divided by all individuals predicted as opioid overdose. PPV is influenced by the prevalence of the outcome of interest.
Negative predictive value (NPV)	The proportion of actual non-overdose cases divided by all observations predicted as non-overdose. When the outcome is rare, NPV is typically high.
Positive likelihood ratio (PLR)	The probability that a person with an actual incident opioid overdose is predicted as opioid overdose, divided by the probability of a person who did not have an actual incident opioid overdose is predicted as opioid overdose. The larger the PLR (>1), the better the prediction performance of an algorithm.
Negative likelihood ratio (NLR)	The probability that a person with an actual incident opioid overdose is predicted as non-overdose, divided by the probability that a person who did not have an actual opioid overdose is predicted as non-overdose. The smaller the NLR (i.e., closer to 0), the better the prediction performance.
Overall misclassification rate	The proportion of incorrectly predicted observations (i.e., false positives and false negatives of opioid overdose) divided by the total number of observations.
F1 score	The weighted average of precision (or PPV) and recall (or sensitivity). F1 takes both false positives and false negatives into account, and it is usually more useful than the overall misclassification rate under an uneven class distribution (e.g., non-overdose individuals comprised the majority of the cohort).(35) An F1 closer to 1 is desirable.
F2 score	The F2 score is the weighted harmonic mean of the precision and recall. Unlike the F1 score, which gives equal weight to precision and recall, the F2 score gives more weight to recall (penalizing the model more for false negatives then false positives).(36) An F2 closer to 1 is desirable.
C-statistic	The area under the receiver operating characteristics curve (ROC) curve, which is a plot of sensitivity vs. (1-specificity) for all potential cut-off probability thresholds for an algorithm. Comparisons of C-statistics based on imbalanced data or rare outcomes can be misleading because C-statistics do not incorporate information about the prevalence or pre-test probability of the outcome.(29)
Precision-recall curves	A precision-recall curve of precision (or PPV; y-axis) vs. recall (sensitivity; x-axis). The curve closer to the upper right corner (corresponding to 100% precision and 100% recall) has better performance.

Number needed to evaluate (NNE)	The NNE is the number of patients necessary to evaluate or screen to detect one outcome (i.e., overdose), similar to the number needed to treat in clinical trials. A PPV of 10% is equivalent to an NNE of 10.
Estimated rate of alerts	Provides the estimated number of alerts per number of patients screened or evaluated over a period of time—for example, per 100 patient over 30 days or 3 months. Too many alerts may lead to alert fatigue; too few may lead to unfamiliarity with the clinical response.

eFigure 5. Performance matrix for predicting opioid overdose using gradient boosting machine (GBM): 2013-2016 internal validation Pennsylvania Medicaid (blue), and 2017-2018 external validation Pennsylvania (orange), and 2015-2017 Arizona Medicaid (green) data: sensitivity analyses using patient-level data



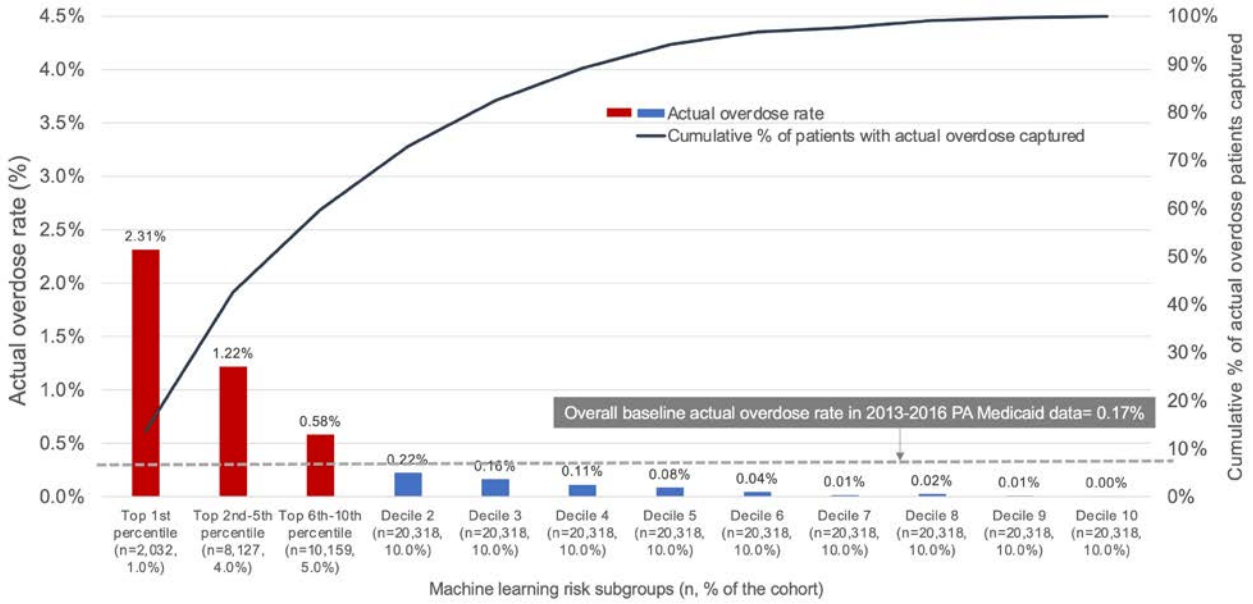
To ensure the GBM algorithm derived from the episode-level data (i.e., beneficiaries may have multiple 3-month periods until occurrence of a censored event including disenrollment or death) can perform well using patient-level data, we iteratively and randomly selected patient-level data (i.e., selecting one 3-month period for each beneficiary) to validate our GBM model. The above figure shows four prediction performance matrices using an example of *patient-level* data from the internal and external validation samples (2013-2016 PA internal validation data: 213,231 beneficiaries with 212,888 non-overdose patients and 343 overdose patients; 2017-2018 PA external validation data: 318,585 beneficiaries with 317,673 non-overdose patients and 912 overdose patients; 2015-2017 AZ external validation data: 391,959 beneficiaries with 391,617 non-overdose patients and 342 overdose patients). **eFigure 4A** shows the areas under the ROC curves (or C-statistics); **eFigure 4B** shows the precision-recall curves (precision=PPV and recall=sensitivity) - precision recall curves that are closer to the upper right corner or above the other method have improved performance; **eFigure 4C** shows the number needed to evaluate by different cutoffs of sensitivity; and **eFigure 4D** shows alerts per 100 patients by different cutoffs of sensitivity.

Abbreviations: AUC: area under the curves; **AZ:** Arizona; **PA:** Pennsylvania; **ROC:** receiver operating characteristics.

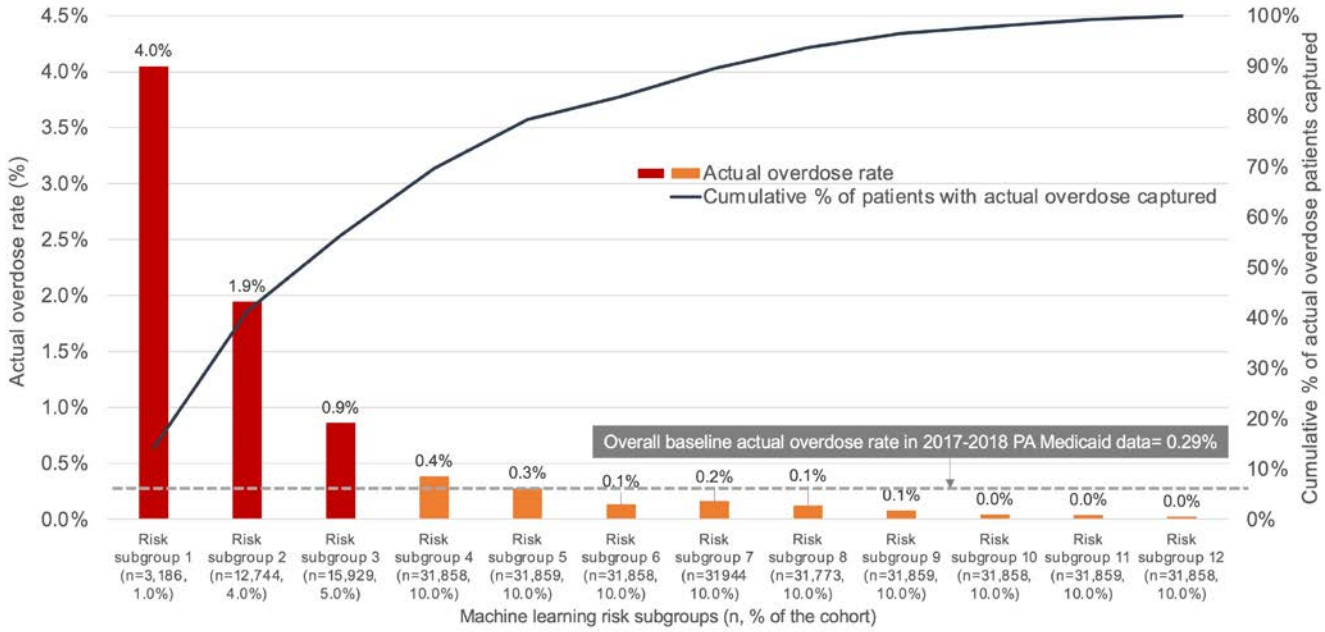
Lo-Ciganic et al. Using machine learning to predict opioid overdose in Medicaid

eFigure 6. Opioid overdose episodes identified by risk subgroup in the 2016-2017 internal-validation Pennsylvania, 2017-2018 external validation Pennsylvania, and 2015-2017 external validation Arizona Medicaid data using gradient boosting machine (GBM): *Using risk score thresholds identified from each validation sample*

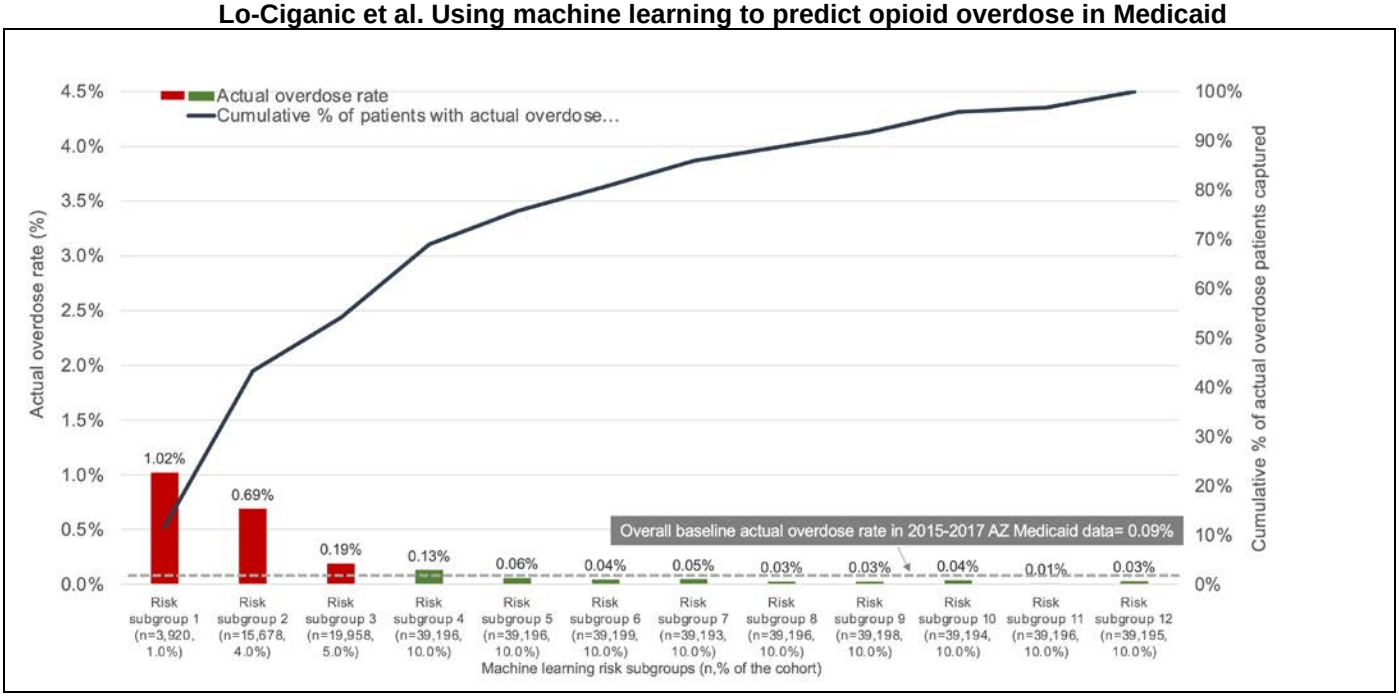
A. 2013-2016 Internal-Validation Pennsylvania Data



B. 2017-2018 External-Validation Pennsylvania Data



C. 2015-2017 External-Validation Arizona Data



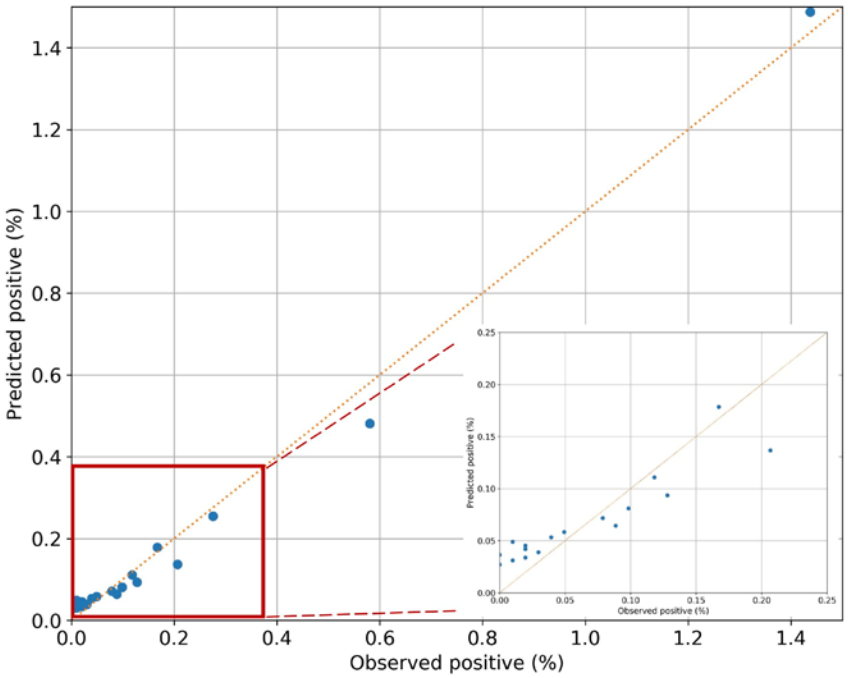
[a]: Based on the individual's predicted probability of an opioid overdose (fatal/nonfatal) event, we classified beneficiaries in the validation samples into risk subgroups. As an alternate risk stratification method, we conducted secondary analyses using the decile risk score thresholds derived from each validation data set (i.e., thresholds varied by validation data) to stratify beneficiaries into 12 risk subgroups (decile groups, with the highest risk decile further split into 3 additional strata based on the top 1, 2nd to 5th, and 6th to 10th percentiles to allow closer examination of patients at highest risk of experiencing overdose). The thresholds of the risk scores derived from the 2013-2016 Pennsylvania internal-validation sample to identify a beneficiary's risk subgroup are as follows: top 1st percentile (≥98.4); 2nd-5th percentile (85.0 ≤risk score <98.4); 6th-10th percentile (63.6 ≤risk score <85.0); decile 2 (47.8 ≤risk score <63.6); decile 3 (38.7 ≤risk score <47.8); decile 4 (32.5 ≤risk score <38.7); decile 5 (27.8 ≤risk score <32.5); decile 6 (23.9 ≤risk score <27.8); decile 7 (20.4 ≤risk score <23.9); decile 8 (17.3 ≤risk score <20.4); decile 9 (14.1 ≤risk score <17.3); decile 10 (14.1 <risk score). The thresholds of the risk scores derived from the 2017-2018 Pennsylvania external-validation sample to identify a beneficiary's risk subgroup are as follows: top 1st percentile (≥ 97.8); 2nd-5th percentile (94.9 ≤risk score <97.8); 6th-10th percentile (94.9≤ risk score <64.0); decile 2 (49.9 ≤risk score <64.0); decile 3 (41.7 ≤risk score <49.9); decile 4 (35.7 ≤risk score <41.7); decile 5 (30.9 ≤risk score <35.7); decile 6 (26.8 ≤risk score <30.9); decile 7 (23.0 ≤risk score <26.8); decile 8 (19.4 ≤risk score <23.0); decile 9 (15.5 ≤risk score <19.4); decile 10 (15.5 <risk score). The thresholds of the risk scores derived from the 2015-2017 Arizona external-validation sample to identify a beneficiary's risk subgroup are as follows: top 1st percentile (≥ 97.4); 2nd-5th percentile (62.8 ≤risk score <97.4); 6th-10th percentile (48.0≤ risk score <62.8); decile 2 (35.6 ≤risk score <48.0); decile 3 (29.3 ≤risk score <35.6); decile 4 (25.1 ≤risk score <29.3); decile 5 (22.0 ≤risk score <25.1); decile 6 (19.4 ≤risk score <22.0); decile 7 (17.2 ≤risk score <19.4); decile 8 (15.0 ≤risk score <17.2); decile 9 (12.7 ≤risk score <15.0); decile 10 (12.7 <risk score).

Lo-Ciganic et al. Using machine learning to predict opioid overdose in Medicaid

eFigure 7. Calibration plots for the 2016-2017 internal-validation Pennsylvania, 2017-2018 external validation Pennsylvania, and 2015-2017 external validation Arizona Medicaid data using gradient boosting machine (GBM)

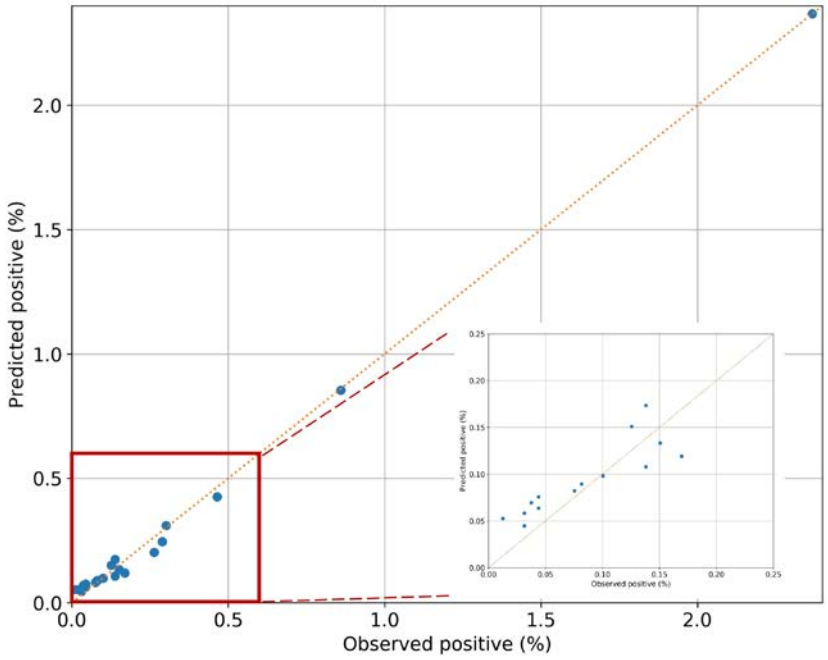
A1. 2013-2016 Internal-Validation Pennsylvania Data (for 20 population bins of equal size)

$$Calibration\ function: predicted\ probability = \frac{e^{-8.68+4.61\times predicted\ score}}{1+e^{-8.68+4.61\times predicted\ score}}$$



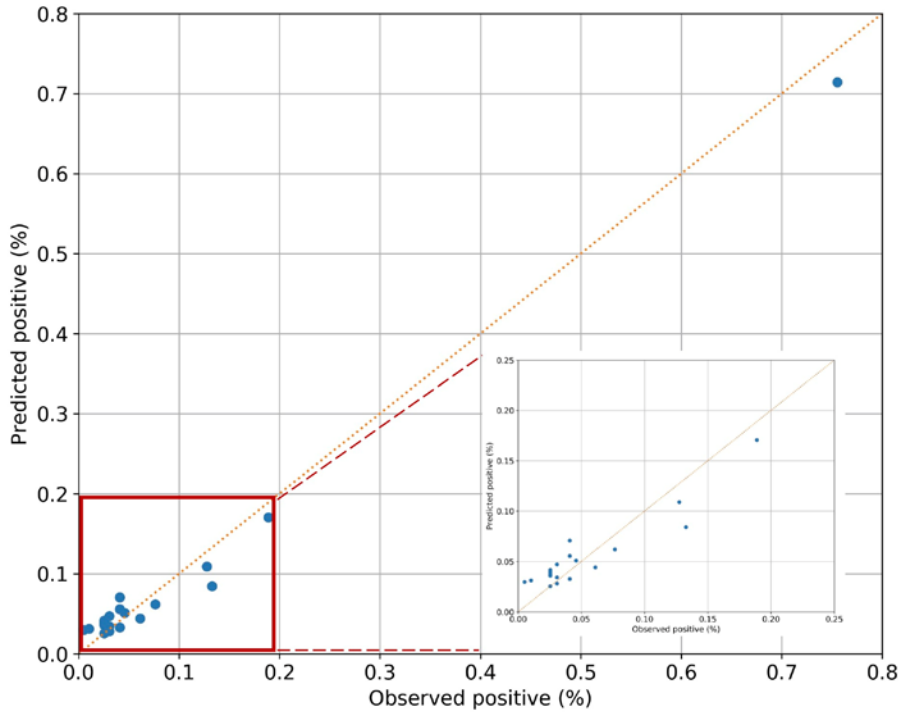
B. 2017-2018 External-Validation Pennsylvania Data (20 bins)

$$Calibration\ function: predicted\ probability = \frac{e^{-8.21+4.63\times predicted\ score}}{1+e^{-8.21+4.63\times predicted\ score}}$$

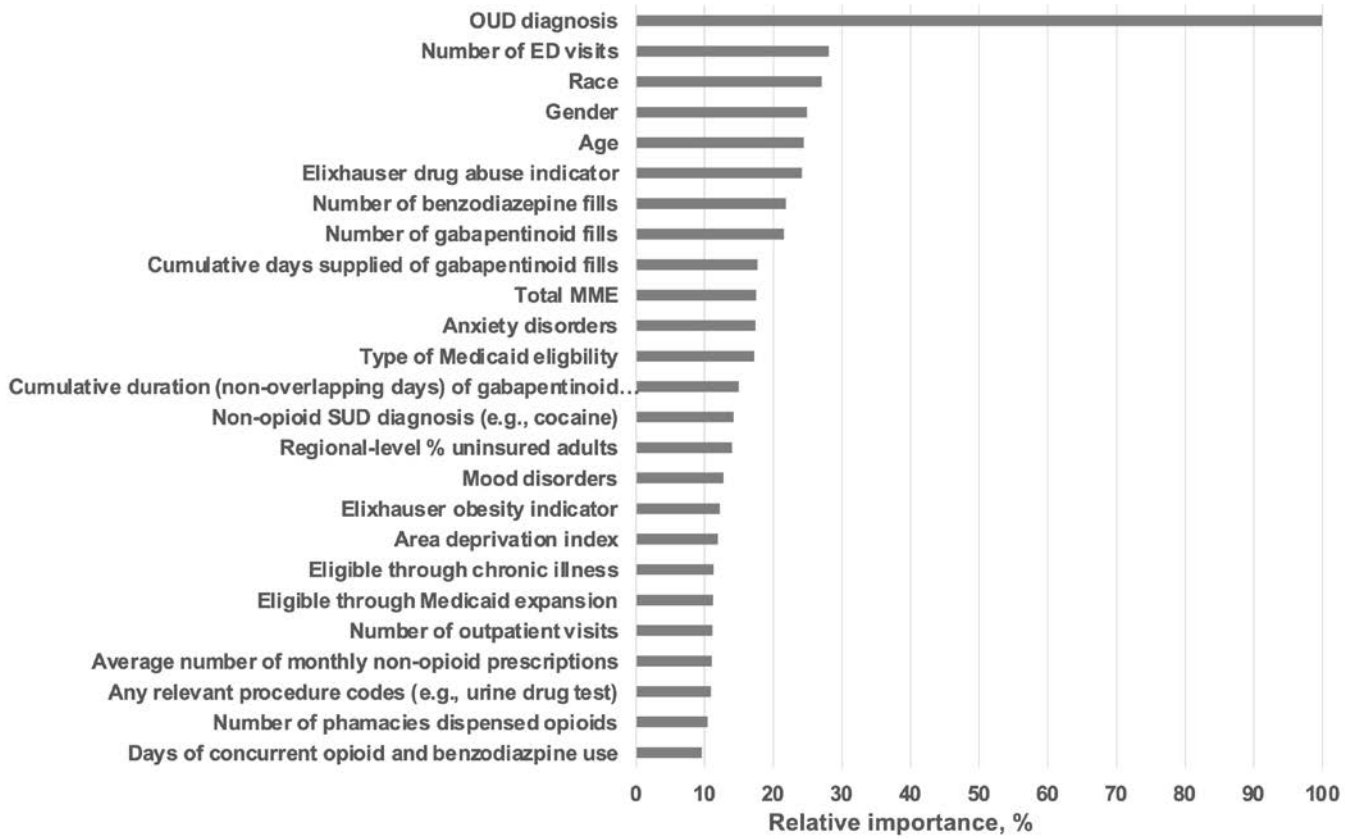


C. 2015-2017 External-Validation Arizona Data (20 bins)

$$Calibration\ function: predicted\ probability = \frac{e^{-8.68+4.23\times predicted\ score}}{1+e^{-8.68+4.23\times predicted\ score}}$$



eFigure 8. Top 25 important predictors for opioid overdose in 2013-2016 Pennsylvania Medicaid data selected by gradient boosting machine[a]



[a]Rather than p values or coefficients, the GBM reports the importance of predictor variables included in a model. Importance is a measure of each variable's cumulative contribution toward reducing square error, or heterogeneity within the subset, after the data set is sequentially split based on that variable. Thus, it reflects a variable's impact on the predictor. Absolute importance is then scaled to give relative importance, with a maximum importance of 100. Among 117 important predictors identified from GBM, the top 10 important predictors included having a diagnosis of OUD, total number of ED visits, race, gender, age, having a diagnosis of drug abuse in the Elixhauser index, total numbers of benzodiazepine fills (e.g., >3), total number of gabapentinoid fills, cumulative days of supply of gabapentinoid use (e.g., >35 days), and total MME.

Abbreviations: ED: emergency department; **GBM**: gradient boosting machine; **MME**: morphine milligram equivalent; **OUD**: opioid use disorder.

eFigure 9. Performance matrix for predicting *fatal* opioid overdose using gradient boosting machine (GBM): 2015-2017 Arizona external validation Medicaid data

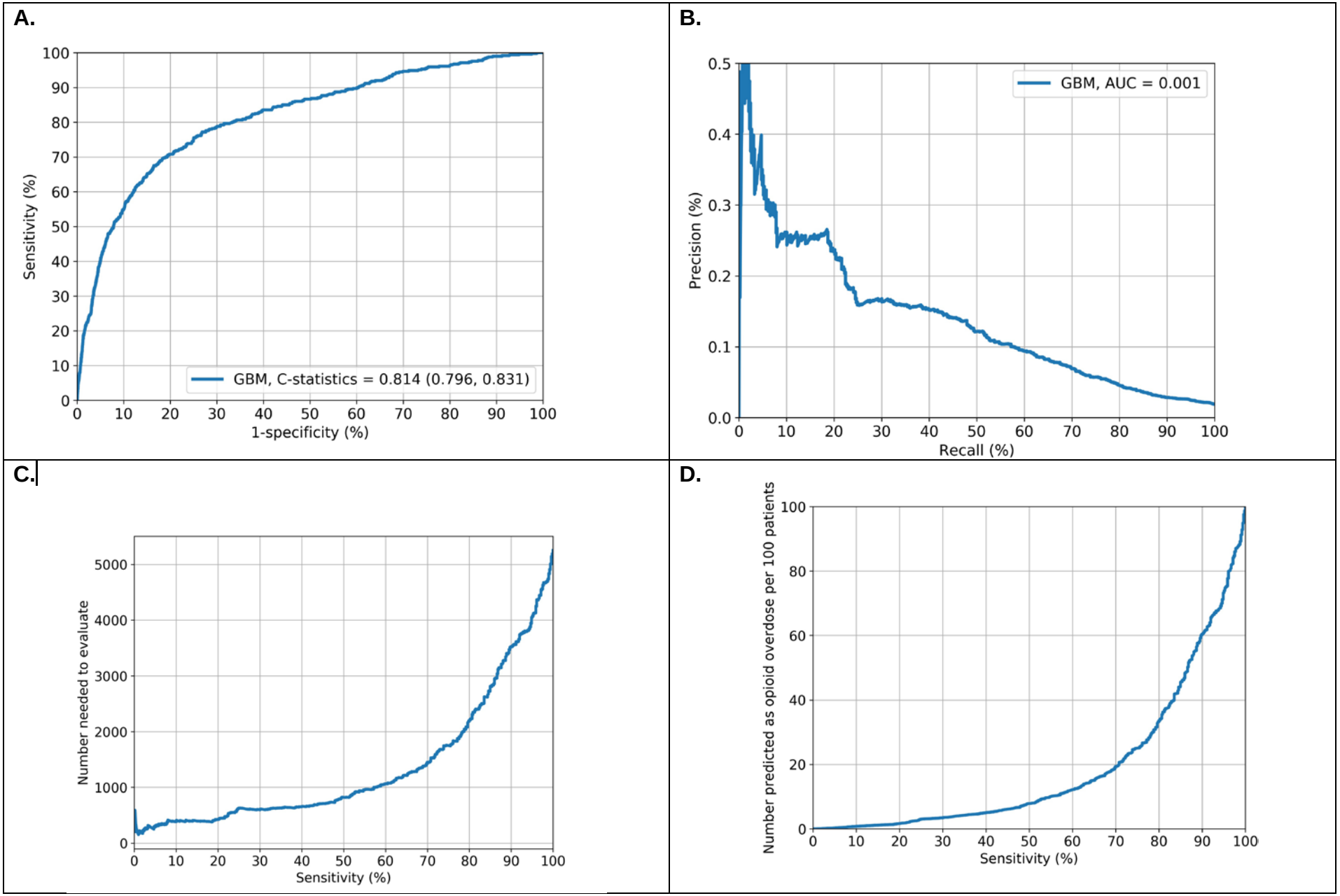


Figure 9. Legend:

eFigure 9 shows four prediction performance matrices in the 2015-2017 Arizona external validation samples (391,959 beneficiaries with 2,550,725 non-overdose episodes and 486 overdose episodes). **eFigure 9A** shows the areas under the ROC curves (or C-statistics); **eFigure 9B** shows the precision-recall curves (precision=PPV and recall=sensitivity): precision recall curves that are closer to the upper right corner or are above another method have improved performance; **eFigure 9C** shows the number needed to evaluate by different cutoffs of sensitivity; and **eFigure 9D** shows alerts per 100 patients by different cutoffs of sensitivity.

Abbreviations: AUC: area under the curves; **AZ:** Arizona; **PA:** Pennsylvania; **ROC:** receiver operating characteristics.

References

1. Webster LR, Webster RM. Predicting aberrant behaviors in opioid-treated patients: preliminary validation of the Opioid Risk Tool. Pain Med. 2005;6(6):432-42.
2. Ives TJ, Chelminski PR, Hammett-Stabler CA, Malone RM, Perhac JS, Potisek NM, et al. Predictors of opioid misuse in patients with chronic pain: a prospective cohort study. BMC Health Serv Res. 2006;6:46.
3. Becker WC, Sullivan LE, Tetrault JM, Desai RA, Fiellin DA. Non-medical use, abuse and dependence on prescription opioids among U.S. adults: Psychiatric, medical and substance use correlates. Drug and Alcohol Dependence. 2008;94(1):38-47.
4. Hall AJ, Logan JE, Toblin RL, et al. Patterns of abuse among unintentional pharmaceutical overdose fatalities. JAMA. 2008;300(22):2613-20.
5. CDC. Overdose deaths involving prescription opioids among Medicaid enrollees - Washington, 2004-2007. MMWR. 2009;58(42):1171-5.
6. White AG, Birnbaum HG, Schiller M, Tang J, Katz NP. Analytic models to identify patients at risk for prescription opioid abuse. Am J Manag Care. 2009;15(12):897-906.
7. Dunn KM, Saunders KW, Rutter CM, Banta-Green CJ, Merrill JO, Sullivan MD, et al. Opioid prescriptions for chronic pain and overdose: a cohort study. Ann Intern Med. 2010;152(2):85-92.
8. Edlund MJ, Martin BC, Devries A, Fan MY, Braden JB, Sullivan MD. Trends in use of opioids for chronic noncancer pain among individuals with mental health and substance use disorders: the TROUP study. The Clinical journal of pain. 2010;26(1):1-8.
9. Sullivan MD, Edlund MJ, Fan MY, Devries A, Brennan Braden J, Martin BC. Risks for possible and probable opioid misuse among recipients of chronic opioid therapy in commercial and Medicaid insurance plans: The TROUP Study. Pain. 2010;150(2):332-9.
10. Bohnert AS, Valenstein M, Bair MJ, Ganoczy D, McCarthy JF, Ilgen MA, et al. Association between opioid prescribing patterns and opioid overdose-related deaths. JAMA. 2011;305(13):1315-21.
11. Volkow ND, McLellan TA, Cotto JH, Karithanom M, Weiss SR. Characteristics of opioid prescriptions in 2009. JAMA. 2011;305(13):1299-301.
12. Webster LR, Cochella S, Dasgupta N, Fakata KL, Fine PG, Fishman SM, et al. An analysis of the root causes for opioid-related overdose deaths in the United States. Pain Medicine (Malden, Mass). 2011;12 Suppl 2:S26-S35.
13. Cepeda MS, Fife D, Chow W, Mastrogiovanni G, Henderson SC. Assessing opioid shopping behaviour: a large cohort study from a medication dispensing database in the US. Drug Safety. 2012;35(4):325-34.
14. Peirce GL, Smith MJ, Abate MA, Halverson J. Doctor and pharmacy shopping for controlled substances. Medical Care. 2012;50(6):494-500.
15. Rice JB, White AG, Birnbaum HG, Schiller M, Brown DA, Roland CL. A Model to Identify Patients at Risk for Prescription Opioid Abuse, Dependence, and Misuse. Pain Medicine. 2012;13(9):1162-73.
16. Baumblatt JA, Wiedeman C, Dunn JR, Schaffner W, Paulozzi LJ, Jones TF. High-Risk Use by Patients Prescribed Opioids for Pain and Its Role in Overdose Deaths. JAMA internal medicine. 2014.
17. Hylan TR, Von Korff M, Saunders K, Masters E, Palmer RE, Carrell D, et al. Automated prediction of risk for problem opioid use in a primary care setting. J Pain. 2015;16(4):380-7.
18. Zedler B, Xie L, Wang L, Joyce A, Vick C, Brigham J, et al. Development of a Risk Index for Serious Prescription Opioid-Induced Respiratory Depression or Overdose in Veterans' Health Administration Patients. Pain Med. 2015;16(8):1566-79.
19. Cochran G, Gordon AJ, Lo-Ciganic WH, Gellad WF, Frazier W, Lobo C, et al. An Examination of Claims-based Predictors of Overdose from a Large Medicaid Program. Med Care. 2017;55(3):291-8.
20. Carey CM, Jena AB, Barnett ML. Patterns of Potential Opioid Misuse and Subsequent Adverse Outcomes in Medicare, 2008 to 2012. Ann Intern Med. 2018;168(12):837-45.
21. Glanz JM, Narwaney KJ, Mueller SR, Gardner EM, Calcaterra SL, Xu S, et al. Prediction Model for Two-Year Risk of Opioid Overdose Among Patients Prescribed Chronic Opioid Therapy. J Gen Intern Med. 2018.

22. Rose AJ, Bernson D, Chui KKH, Land T, Walley AY, LaRochelle MR, et al. Potentially Inappropriate Opioid Prescribing, Overdose, and Mortality in Massachusetts, 2011-2015. J Gen Intern Med. 2018;33(9):1512-9.
23. Zedler BK, Saunders WB, Joyce AR, Vick CC, Murrelle EL. Validation of a Screening Risk Index for Serious Prescription Opioid-Induced Respiratory Depression or Overdose in a US Commercial Health Plan Claims Database. Pain Med. 2018;19(1):68-78.
24. Hacker K, Jones LD, Brink L, Wilson A, Cherna M, Dalton E, et al. Linking Opioid-Overdose Data to Human Services and Criminal Justice Data: Opportunities for Intervention. Public Health Rep. 2018;133(6):658-66.
25. Lo-Ciganic WH, Huang JL, Zhang HH, Weiss JC, Wu Y, Kwoh CK, et al. Evaluation of Machine-Learning Algorithms for Predicting Opioid Overdose Risk Among Medicare Beneficiaries With Opioid Prescriptions. JAMA Netw Open. 2019;2(3):e190968.
26. County Health Rankings and Roadmaps: Building a Culture of Health, County by County [Available from: http://www.countyhealthrankings.org/explore-health-rankings/use-data. Accessed: May 4, 2019.
27. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): the TRIPOD statement. Ann Intern Med. 2015;162(1):55-63.
28. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig L, et al. STARD 2015: an updated list of essential items for reporting diagnostic accuracy studies. BMJ. 2015;351:h5527.
29. Romero-Brufau S, Huddleston JM, Escobar GJ, Liebow M. Why the C-statistic is not informative to evaluate early warning scores and what metrics to use. Crit Care. 2015;19:285.
30. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J. 2005;47(4):458-72.
31. Friedman JH. Greedy Function Approximation: A Gradient Boosting Machine. Technical report, Dept. of Statistics, Stanford University. 1999.
32. Friedman JH. A Gradient Boosting Machine. Ann Stat. 2001;29(5):1189.
33. Gellad WF, Thorpe JM, Zhao X, Thorpe CT, Sileanu FE, Cashy JP, et al. Impact of Dual Use of Department of Veterans Affairs and Medicare Part D Drug Benefits on Potentially Unsafe Opioid Use. Am J Public Health. 2018;108(2):248-55.
34. Bohnert ASB, Valenstein M, Bair MJ, Ganoczy D, McCarthy JF, Ilgen MA, et al. Association between opioid prescribing patterns and opioid overdose-related deaths. JAMA. 2011;305(13):1315-21.
35. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
36. Siblini W, Fréry J, He-Guelton L, Oblé F, Wang Y-Q, editors. Master Your Metrics with Calibration2020; Cham: Springer International Publishing.

Appendix A. Compliance to the 2015 Standards for Reporting of Diagnostic Accuracy (STARD) Checklist

Section & Topic	No	Item	Reported on page #
TITLE OR ABSTRACT			
	1	Identification as a study of diagnostic accuracy using at least one measure of accuracy (such as sensitivity, specificity, predictive values, or AUC)	4-5
ABSTRACT	**2**	Structured summary of study design, methods, results, and conclusions (for specific guidance, see STARD for Abstracts)	4-5
INTRODUCTION			
	3	Scientific and clinical background, including the intended use and clinical role of the index test	6
	4	Study objectives and hypotheses	6
METHODS			
Study design	**5**	Whether data collection was planned before the index test and reference standard were performed (prospective study) or after (retrospective study)	7
Participants	**6**	Eligibility criteria	8-9
	7	On what basis potentially eligible participants were identified (such as symptoms, results from previous tests, inclusion in registry)	7-8
	8	Where and when potentially eligible participants were identified (setting, location and dates)	7-8
	9	Whether participants formed a consecutive, random or convenience series	7
Test methods	**10a**	Index test, in sufficient detail to allow replication	9-11; Appendix Methods
	10b	Reference standard, in sufficient detail to allow replication	9; eTable 2
	11	Rationale for choosing the reference standard (if alternatives exist)	N/A
	12a	Definition of and rationale for test positivity cut-offs or result categories of the index test, distinguishing pre-specified from exploratory	13-14; eTable 3
	12b	Definition of and rationale for test positivity cut-offs or result categories of the reference standard, distinguishing pre-specified from exploratory	9; eTable 2
	13a	Whether clinical information and reference standard results were available to the performers/readers of the index test	9; eTable 2
	13b	Whether clinical information and index test results were available to the assessors of the reference standard	8-9
Analysis	**14**	Methods for estimating or comparing measures of diagnostic accuracy	10
	15	How indeterminate index test or reference standard results were handled	10
	16	How missing data on the index test and reference standard were handled	N/A
	17	Any analyses of variability in diagnostic accuracy, distinguishing pre-specified from exploratory	12
	18	Intended sample size and how it was determined	8
RESULTS			
Participants	**19**	Flow of participants, using a diagram	eFigures 1-2
	20	Baseline demographic and clinical characteristics of participants	13; Table 1
	21a	Distribution of severity of disease in those with the target condition	13; Table 1
	21b	Distribution of alternative diagnoses in those without the target condition	N/A
	22	Time interval and any clinical interventions between index test and reference standard	N/A
Test results	**23**	Cross tabulation of the index test results (or their distribution) by the results of the reference standard	13
	24	Estimates of diagnostic accuracy and their precision (such as 95% confidence intervals)	eTable 4
	25	Any adverse events from performing the index test or the reference standard	N/A
DISCUSSION			
	26	Study limitations, including sources of potential bias, statistical uncertainty, and generalisability	17-18
	27	Implications for practice, including the intended use and clinical role of the index test	17

	OTHER INFORMATION			
		28	Registration number and name of registry	N/A
		29	Where the full study protocol can be accessed	N/A
		30	Sources of funding and other support; role of funders	19

Appendix B. Compliance to the 2015 Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis Or Diagnosis (TRIPOD) Checklist

Section/Topic	Item		Checklist Item	Page
Title and abstract				
Title	1	D;V	Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted.	1
Abstract	2	D;V	Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions.	4-5
Introduction				
Background and objectives	3a	D;V	Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models.	6
	3b	D;V	Specify the objectives, including whether the study describes the development or validation of the model or both.	6
Methods				
Source of data	4a	D;V	Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable.	7
	4b	D;V	Specify the key study dates, including start of accrual; end of accrual; and, if applicable, end of follow-up.	7
Participants	5a	D;V	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including number and location of centres.	7
	5b	D;V	Describe eligibility criteria for participants.	8-9
	5c	D;V	Give details of treatments received, if relevant.	8-9
Outcome	6a	D;V	Clearly define the outcome that is predicted by the prediction model, including how and when assessed.	8-9
	6b	D;V	Report any actions to blind assessment of the outcome to be predicted.	NA
Predictors	7a	D;V	Clearly define all predictors used in developing or validating the multivariable prediction model, including how and when they were measured.	9
	7b	D;V	Report any actions to blind assessment of predictors for the outcome and other predictors.	NA
Sample size	8	D;V	Explain how the study size was arrived at.	eFigure 1
Missing data	9	D;V	Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.	Appendix Methods
Statistical analysis methods	10a	D	Describe how predictors were handled in the analyses.	9-10; Appendix Methods
	10b	D	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.	9-11; Appendix Methods
	10c	V	For validation, describe how the predictions were calculated.	10-11; Appendix Methods
	10d	D;V	Specify all measures used to assess model performance and, if relevant, to compare multiple models.	10; eFigure3
	10e	V	Describe any model updating (e.g., recalibration) arising from the validation, if done.	12
Risk groups	11	D;V	Provide details on how risk groups were created, if done.	11
Development vs. validation	12	V	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.	10-11; Table 1

Results				
Participants	13a	D;V	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	9; eFigure 1
	13b	D;V	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.	13; Table 1
	13c	V	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).	Table 1
Model development	14a	D	Specify the number of participants and outcome events in each analysis.	13; Table 1
	14b	D	If done, report the unadjusted association between each candidate predictor and outcome.	N/A
Model specification	15a	D	Present the full prediction model to allow predictions for individuals (i.e., all regression coefficients, and model intercept or baseline survival at a given time point).	Consult investigators
	15b	D	Explain how to the use the prediction model.	Appendix Methods
Model performance	16	D;V	Report performance measures (with CIs) for the prediction model.	13-14; eTable 3
Model-updating	17	V	If done, report the results from any model updating (i.e., model specification, model performance).	9; eFigure 2
Discussion				
Limitations	18	D;V	Discuss any limitations of the study (such as nonrepresentative sample, few events per predictor, missing data).	17-18
Interpretation	19a	V	For validation, discuss the results with reference to performance in the development data, and any other validation data.	17
	19b	D;V	Give an overall interpretation of the results, considering objectives, limitations, results from similar studies, and other relevant evidence.	17-18
Implications	20	D;V	Discuss the potential clinical use of the model and implications for future research.	17
Other information				
Supplementary information	21	D;V	Provide information about the availability of supplementary resources, such as study protocol, Web calculator, and data sets.	Online supplement
Funding	22	D;V	Give the source of funding and the role of the funders for the present study.	19

NA: not applicable

*Items relevant only to the development of a prediction model are denoted by D, items relating solely to a validation of a prediction model are denoted by V, and items relating to both are denoted D;V. We recommend using the TRIPOD Checklist in conjunction with the TRIPOD Explanation and Elaboration document.