Supplemental Materials Accompanying the Manuscript "A Spatio-temporal Cluster-aware Supervised Learning Framework for Predicting County-level Drug Overdose Deaths"

Table S1. Variables included in the county-level prediction model for drug overdose deaths.

	Description	Data sources
Prediction outcome		
Drug overdose deaths	Annual number of overdose deaths from any drugs, which are identified by International Classification of Diseases Codes 10th Revision (ICD-10 codes) X40-44, X60-64, X85, and Y10-14 for the underlying cause of death and ICD-10 codes T36-50 for multiple causes of death.	Centers for Disease Control and Prevention (CDC) Wide-ranging ONline Data for Epidemiologic Research (WONDER) database [1]
Predictor variables		
Supply-side factors		
Drug-related crime incidence	Annual number of crime incidents that involve at least one of the 18 documented drug types aggregated at the county level. For agencies covering multiple counties, we distributed the crime incidence across the counties, assuming it was distributed evenly based on population size.	National Incidence-Based Reporting System (NIBRS) [2]
Percentage of fentanyl among drug seizures	Percentage of fentanyl detected in drug seizures in each year by state.	National Forensic Laboratory Information System (NFLIS) [3]
Health factors		
Number of primary care physicians	Number of primary care physicians in a county. Primary care physicians include practicing non-federal physicians (M.D.s and D.O.s) under age 75 specializing in general practice medicine, family medicine, internal medicine, and pediatrics.	County Health Rankings & Roadmap [4]
Annual poor mental health days	Average (age-adjusted) number of mentally unhealthy days reported in the past 30 days, based on responses to The Behavioral Risk Factor Surveillance System (BRFSS) question: "Now thinking about your mental health, which includes stress, depression, and problems with emotions, for how many days during the past 30 days was your mental health not good?"	County Health Rankings & Roadmap [4]
Annual poor physical health days	Average (age-adjusted) number of physically unhealthy days reported in the past 30 days, based on responses to The Behavioral Risk Factor Surveillance System (BRFSS) question: "Now thinking about your physical health, which includes physical illness and injury, for how many	County Health Rankings & Roadmap [4]

	days during the past 30 days was your physical health not good?"	
Adult smoking rate	Age-adjusted percentage of the adult population in a county who both report that they currently smoke every day or some days and have smoked at least 100 cigarettes in their lifetime.	County Health Rankings & Roadmap [4]
Adult obesity rate	Age-adjusted percentage of the adult population (ages 18 and older) that reports a body mass index (BMI) greater than or equal to 30 kg/m^2 in BRFSS (participants are asked to self-report their height and weight; BMIs are calculated from these reported values).	County Health Rankings & Roadmap [4]
Socioeconomic factors		
Population size	The total number of people residing in the county	American Community Survey [5]
Household median income	Estimated median of the annual household income of the county	American Community Survey [5]
High school graduation rate	Percentage of people living in the county estimated to have graduated from high school or received an equivalent certification	American Community Survey [5]
Uninsurance rate	Percentage of adults who are living without health insurance coverage	American Community Survey [5]
Unemployment rate	Percentage of people in the labor force who are unemployed	American Community Survey [5]
Gini index	A continuous measure of income inequality that summarizes the county's income distribution	American Community Survey [5]
Rural-Urban Commuting Area (RUCA) code	A nine-category variable created by the US Department of Agriculture (USDA)'s Economic Research Service using measures of population density, urbanization, and daily commuting, coded in a spectrum with 1 being the most metropolitan area and 9 being the least	USDA RUCA code [6]
Geospatial factors		
Geographic coordinates	Longitudinal and latitude of the centroid of the county	Google Distance Matrix API [7]
State border indicator	Binary indicator with 1 representing that the county is bordering another US state.	Extracted from map
Highway indicator	Binary indicator, which equals 1 if at least one interstate highway passes through the county.	Extracted from map
Driving distance between counties	Driving distance between the centroids of two counties, which are used to drive the drug-related "crime gravity" measure. Crime gravity is defined as the total drug-related crime rates in all counties within a radius of 50 miles of the given county, weighted by the inverse of the squared driving distance from these counties.	Google Distance Matrix API [7]

Table S2. Combinations of hyperparameters of CASL model for model selection.

Hyperparameters	**Values**
Number of clusters, K	1, 2, 3, 4, 5
Weight parameter for suppressed values, κ	0.1, 1, 10, 100
Weight parameter for clustering performance, ρ	0.1, 0.2, 0.5, 1, 5, 10
Cluster-specific regularization coefficient for the supervised learning model, $\boldsymbol{\lambda} = \{\lambda_k\}_{k\in[K]}$	
K = 1	$\lambda = 0.1, 0.2, 0.5, 1, 10$
K = 2	$\{(\lambda_1, \lambda_2) \mid \lambda_1 = \lambda_2 \text{ OR } \lambda_1 < \lambda_2; \lambda_1, \lambda_2 = 0.1, 0.2, 0.5, 1, 5, 10\}$
K = 3	$\{(\lambda_1, \lambda_2, \lambda_3) \mid \lambda_1 = \lambda_2 = \lambda_3 \text{ OR } \lambda_1 < \lambda_2 < \lambda_3;\ \lambda_1, \lambda_2, \lambda_3 = 0.1, 0.2, 0.5, 1, 5, 10\}$
K = 4	$\{(\lambda_1, \lambda_2, \lambda_3, \lambda_4) \mid \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 \text{ OR } \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4;\ \lambda_1, \lambda_2, \lambda_3, \lambda_4 = 0.1, 0.2, 0.5, 1, 5, 10\}$
K = 5	$\{(\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) \mid \lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = \lambda_5 \text{ OR } \lambda_1 < \lambda_2 < \lambda_3 < \lambda_4 < \lambda_5;\ \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5 = 0.1, 0.2, 0.5, 1, 5, 10\}$

Table S3. Hyperparameters of random forest and lasso regression for model selection.

Model	**Hyperparameter**	**Values**
Random forest	Number of trees	100, 200, 500, 1000
	Maximum depths	10, 20, 50, 100, unlimited
Lasso regression	L_1 norm regularization parameter	0.1, 0.2, 0.3, 0.4, 0.5, 1, 2, 5, 10, 100

Table S4. Performance comparison in cross-validation and testing of the CASL, random forest, and lasso regression models for the prediction years 2020 and 2021.

Prediction year	Mixing parameter for the mixed error score, ω	Model	Cross-validation			Testing for the prediction year		
			MAE	FPR	Mixed error score	MAE	FPR	Mixed error score
2020	0.1	CASL K=1	12.48	23.95%	6.01	14.54	4.38%	8.92
		Random forest	13.97	8.01%	6.7	16.47	6.25%	10.10
		Lasso regression	13.84	5.76%	6.63	17.06	1.25%	10.47
	1	CASL K=1	12.48	23.95%	6.12	14.54	4.38%	8.94
		Random forest	13.97	8.01%	6.74	16.47	6.25%	10.13
		Lasso regression	13.84	5.76%	6.66	17.06	1.25%	10.47
	10	CASL K=2	12.94	3.81%	6.42	16.97	1.25%	10.46
		Random forest	13.97	8.01%	7.11	16.47	6.25%	10.34
		Lasso regression	13.84	5.76%	6.93	17.06	1.25%	10.52
	100	CASL K=4	13.91	0.82%	7.11	18.36	0.00%	11.27
		Random forest	14.25	6.97%	10.46	16.70	6.25%	12.66
		Lasso regression	14.02	5.01%	9.33	16.86	1.25%	10.83
2021	0.1	CASL K=1	13.49	21.17%	6.94	10.90	12.06%	7.19
		Random forest	15.07	7.28%	7.75	12.12	9.22%	7.99
		Lasso regression	14.91	5.14%	7.69	11.71	11.35%	7.73
	1	CASL K=2	13.51	18.05%	7.03	10.36	12.77%	6.88
		Random forest	15.07	7.28%	7.78	12.12	9.22%	8.02
		Lasso regression	14.91	5.14%	7.71	11.71	11.35%	7.76
	10	CASL K=5	14.25	3.45%	7.51	12.61	11.35%	8.70
		Random forest	15.07	7.28%	8.11	12.12	9.22%	8.30
		Lasso regression	14.91	5.14%	7.95	11.71	11.35%	8.11
	100	CASL K=5	14.95	0.56%	7.98	13.70	0.00%	9.04
		Random forest	15.35	6.52%	11.23	12.56	9.22%	11.42
		Lasso regression	15.09	4.57%	10.08	11.55	11.35%	11.48

Abbreviations: MAE = mean absolute error. FPR = False positive rate.

Table S5. County characteristics of the clusters identified by the CASL model ($K = 4$).

	Prediction for the year 2020				**Prediction for the year 2021**			
County characteristics	**Cluster 1** Mean (SD)	**Cluster 2** Mean (SD)	**Cluster 3** Mean (SD)	**Cluster 4** Mean (SD)	**Cluster 1** Mean (SD)	**Cluster 2** Mean (SD)	**Cluster 3** Mean (SD)	**Cluster 4** Mean (SD)
Population size	964,429.40 (309,144.46)	562,709.27 (202,489.65)	166,757.49 (75,389.60)	33,885.98 (26,881.98)	964,878.53 (310,218.88)	563,391.82 (202,323.51)	177,844.07 (74,162.03)	35,210.53 (28,895.46)
Number of drug overdose deaths								
Past 3-year average	369.27 (96.36)	202.27 (51.55)	59.94 (26.62)	6.33 (9.47)	378.40 (120.82)	219.06 (70.98)	67.13 (26.82)	8.18 (11.26)
Year of prediction	426.60 (165.39)	254.35 (89.57)	73.67 (29.94)	13.05 (10.30)	469.30 (158.45)	285.06 (123.78)	83.94 (34.84)	16.43 (14.56)
Number of drug-related crime incidents								
Past 3-year average	2523.17 (1917.28)	2219.69 (1624.76)	698.79 (649.32)	162.95 (202.50)	2133.70 (1616.17)	2004.59 (1372.61)	719.83 (668.66)	156.56 (193.12)
Year of prediction	1446.10 (566.23)	1649.47 (994.37)	578.33 (627.93)	140.81 (170.06)	1566.80 (699.54)	1537.41 (1229.49)	675.74 (694.15)	156.61 (185.28)
Number of primary care physicians	981.63 (439.47)	526.73 (258.31)	126.95 (89.22)	19.96 (28.59)	982.77 (439.84)	529.47 (260.50)	136.81 (91.83)	20.47 (29.18)
Median household income ($)	67,084.70 (15,215.15)	66,232.55 (16,278.64)	62,137.09 (14,785.10)	47,570.18 (11,482.46)	69,647.87 (15,857.14)	68,260.51 (16,493.65)	65,537.73 (14,976.12)	49,639.93 (11,875.54)
Number of annual poor physical health days	3.54 (0.35)	3.84 (0.51)	3.90 (0.67)	4.67 (0.68)	3.60 (0.41)	3.84 (0.50)	3.85 (0.62)	4.67 (0.66)
Number of annual poor mental health days	4.22 (0.37)	4.67 (0.63)	4.71 (0.52)	5.32 (0.73)	4.31 (0.37)	4.72 (0.63)	4.76 (0.52)	5.44 (0.69)
Adult smoking rate	0.17 (0.03)	0.18 (0.03)	0.19 (0.03)	0.22 (0.03)	0.16 (0.04)	0.17 (0.03)	0.18 (0.03)	0.22 (0.03)
Adult obesity rate	0.28 (0.03)	0.29 (0.04)	0.31 (0.04)	0.34 (0.04)	0.28 (0.03)	0.29 (0.04)	0.31 (0.05)	0.34 (0.04)
Unemployment rate	0.05 (0.01)	0.05 (0.01)	0.05 (0.01)	0.06 (0.02)	0.04 (0.01)	0.04 (0.01)	0.04 (0.01)	0.05 (0.02)
Gini index	0.48 (0.02)	0.48 (0.03)	0.44 (0.02)	0.45 (0.03)	0.48 (0.02)	0.48 (0.03)	0.44 (0.02)	0.45 (0.03)
High school graduation rate	0.83 (0.07)	0.84 (0.07)	0.90 (0.04)	0.91 (0.06)	0.86 (0.06)	0.85 (0.07)	0.91 (0.04)	0.92 (0.05)
Uninsured rate among adults	0.08 (0.03)	0.09 (0.04)	0.10 (0.03)	0.11 (0.04)	0.07 (0.02)	0.08 (0.04)	0.08 (0.03)	0.10 (0.03)

Abbreviations: SD = standard deviation.

References

1. Centers for Disease Control and Prevention, National Center for Health Statistics. National Vital Statistics System, Mortality 1999-2020 on CDC WONDER Online Database, released in 2021. Retrieved from https://wonder.cdc.gov/.
2. Federal Bureau of Investigation. National Incident-Based Reporting System (NIBRS). Retrieved from https://bjs.ojp.gov/national-incident-based-reporting-system-nibrs.
3. Drug Enforcement Administration. National Forensic Laboratory Information System (NFLIS). Retrieved from https://www.nflis.deadiversion.usdoj.gov/.
4. University of Wisconsin Population Health Institute. County Health Rankings & Roadmaps 2024. Retrieved from https://www.countyhealthrankings.org.
5. US Census Bureau. American Community Survey, [2010-2021]. Retrieved from https://www.census.gov/programs-surveys/acs/.
6. US Department of Agriculture, Economic Research Service. Rural-Urban Commuting Area Codes. Retrieved from https://www.ers.usda.gov/data-products/rural-urban-commuting-area-codes/.
7. Google. Distance Matrix API. Retrieved from https://developers.google.com/maps/documentation/distance-matrix/overview.