PROVIDENT: Development and Validation of a Machine Learning Model to Predict Neighborhood-level Overdose Risk in Rhode Island

Bennett Allen,[a] Robert C. Schell,[b] Victoria A. Jent,[a] Maxwell Krieger,[c] Claire Pratty,[c] Benjamin D. Hallowell,[d] William C. Goedel,[c] Melissa Basta,[d] Jesse L. Yedinak,[c] Yu Li,[c] Abigail R. Cartus,[c] Brandon D. L. Marshall,[c] Magdalena Cerdá,[a] Jennifer Ahern,[e] and Daniel B. Neill[f,g,h]

Background: Drug overdose persists as a leading cause of death in the United States, but resources to address it remain limited. As a result, health authorities must consider where to allocate scarce resources within their jurisdictions. Machine learning offers a strategy to identify areas with increased future overdose risk to proactively allocate overdose prevention resources. This modeling study is embedded in a randomized trial to measure the effect of proactive resource allocation on statewide overdose rates in Rhode Island (RI).
Methods: We used statewide data from RI from 2016 to 2020 to develop an ensemble machine learning model predicting neighborhood-level fatal overdose risk. Our ensemble model integrated gradient boosting machine and super learner base models in a moving window framework to make predictions in 6-month intervals. Our performance target, developed a priori with the RI Department of Health, was to identify the 20% of RI neighborhoods containing at least 40% of statewide overdose deaths, including at least one neighborhood per municipality. The model was validated after trial launch.
Results: Our model selected priority neighborhoods capturing 40.2% of statewide overdose deaths during the test periods and 44.1% of statewide overdose deaths during validation periods. Our ensemble outperformed the base models during the test periods and performed comparably to the best-performing base model during the validation periods.
Conclusions: We demonstrated the capacity for machine learning models to predict neighborhood-level fatal overdose risk to a degree of accuracy suitable for practitioners. Jurisdictions may consider predictive modeling as a tool to guide allocation of scarce resources.

Keywords: Gradient boosting machine; Machine learning; Opioid; Overdose; Rhode Island; Super learner

Keywords: machine learning; overdose; opioid; Super Learner; gradient boosting machine; Rhode Island

Drug overdose remains a leading cause of death in the United States (US), with over 105,000 deaths in 2021 alone.[1] In recent years, the introduction of illicitly manufactured synthetic opioids into the US drug markets has markedly increased overdose deaths.[2] Overdose deaths further increased during the SARS-CoV-2 (COVID-19) pandemic, likely due to disruptions in the drug supply and health care services.[3,4] The rapidly accelerating epidemic demands nimble resource allocation tailored to the local profile of the overdose crisis.

Despite increased public-sector resources devoted to overdose response at local, state, and federal levels,[5] population-level health interventions remain constrained by scarce resources.[6] As such, public health authorities must consider how to allocate overdose prevention resources across the jurisdictions they serve. Decisions about where to allocate preventive interventions (e.g., naloxone distribution, street outreach) typically are made using data on historical area-level fatal and nonfatal overdose burden, often absent other factors. However, changing trends in overdose

following the introduction of fentanyl and other potent synthetic opioids into the illicit drug supply complicate these decisions, as past overdose burden may no longer accurately reflect current community overdose risk.[7] Public health and harm reduction practitioners require new methods to proactively identify future community-level overdose risk and inform service delivery. In the present study, we apply such a method that, while not explicitly modeling trajectories in overdose, incorporates historical overdose data to generate forecasts.

PROVIDENT: Forecasting Neighborhood-level Overdose Risk to Guide Public Health Resource Allocation

This modeling study is embedded within the Preventing Overdose using Information and Data from the Environment (PROVIDENT) randomized controlled trial (NCT05096429), which aims to test the effect of allocating overdose prevention resources in Rhode Island (RI) according to machine learning-based prediction of future overdose risk, in comparison to reactive responses guided by traditional surveillance reports.[8] The trial's central hypothesis is that proactive resource allocation based on machine learning model predictions can more effectively reduce drug overdose-related morbidity and mortality in the context of a spatially dynamic epidemic, compared with standard resource allocation approaches using epidemiologic surveillance. The PROVIDENT trial is conducted in partnership with the RI Department of Health (RIDOH), which centralizes overdose prevention resources and infrastructure across RI. As part of an academic–state health department partnership,[9] the trial randomized each of RI's 39 municipalities to an intervention condition, where model predictions are available to identify prioritized neighborhoods with highest risk of future fatal overdose, or to a control condition receiving overdose prevention resources as usual in accordance with the state's strategic plan and based on routine surveillance reporting without targeted prioritization. Details of the PROVIDENT trial design and protocol are available in a prior publication.[8] This study presents the development and internal validation of the predictive model informing the intervention and describes the evaluation criteria used to assess its performance.

The PROVIDENT trial seeks to determine whether machine learning models might enable public health practitioners to leverage available data to predict future community-level overdose risk and proactively allocate resources. Advances in small-area prediction using spatiotemporal machine learning methods indicate that accurate forecasting of future community overdose risk is possible.[10] Such methods may bolster the impact of limited resources by prioritizing both communities with endemically high rates of overdose death and emerging overdose "hotspots." Machine learning methods leverage a broad array of surveillance data as a single, high-dimensional dataset without the need for theory-driven feature selection.[11] With predictive performance as

the goal, machine learning also can facilitate integration of complementary predictive approaches (e.g., spatiotemporal and tree based) through ensemble techniques.[12]

To inform allocation of overdose prevention interventions and to maximize the impact of finite resources in RI, we developed a machine learning tool to forecast future neighborhood-level overdose burden. Our study builds on prior equity-focused work leveraging machine learning to target interventions in resource-limited settings in substance use and HIV.[13–16] We uniquely focus on area-level overdose prevention interventions, partnering directly with public health practitioners. We note that the majority of machine learning work related to overdose has focused on individual-level prediction to inform clinical intervention,[17–21] whereas our study focuses on community-level prediction to inform public health intervention, a novel area for the application of machine learning. Using a variety of public health, social, environmental, and economic data sources widely available to state and local health authorities in the US and across multiple domains for which prior literature has established associations with neighborhood-level overdose mortality,[8] the model predicts future community-level overdose risk in neighborhoods across RI.

METHODS

Study Setting and Period

This modeling study used RI data from 1 January, 2016 to 30 June, 2020. The trial began in November 2021 and will continue through June 2024, updating model predictions every 6 months. Model predictions for municipalities in the intervention condition are shared with RIDOH and local community organizations through a password-protected web portal to inform harm reduction service delivery.

The neighborhood unit was the census block group (CBG), the smallest geographic unit for which RI overdose mortality data and US Census data are available. CBGs correspond to small areas of approximately 600–3000 residents, and prior research has identified them as valid proxies for neighborhoods,[22] which aligns with the study goal of prioritizing public health interventions at neighborhood level. We use the terms "CBG" and "neighborhood" interchangeably.

As of the 2010 census, RI contains 815 CBGs organized into 39 municipalities. CBGs were defined as urban if they have a population density over 2500 persons per square mile and at least 50% of land developed, and otherwise defined as nonurban.[23] We excluded CBGs with special land use designations (e.g., bodies of water, military bases, or airports) for a final sample of 809 populated CBGs, 57.6% of which were urban (Figure). All procedures were approved by Brown University School of Public Health and RIDOH Institutional Review Boards.

Data Sources

This study used five sources of data, aggregated to CBG level: (1) overdose mortality[24]; (2) emergency medical services (EMS)-attended nonfatal opioid overdoses[25]; (3) Prescription
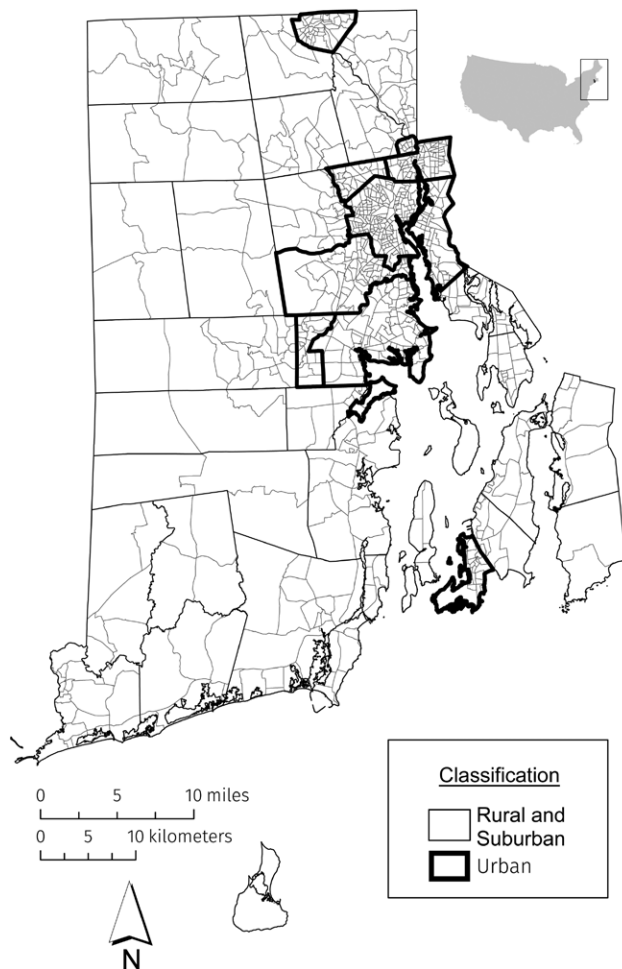
FIGURE. Rhode Island location within the contiguous United States.

Drug Monitoring Program (PDMP) data at both patient and prescription levels[26]; (4) American Community Survey (ACS) data[27]; and (5) public access land use, health care, and social service availability data[28–33] (eAppendix A; http://links.lww.com/EDE/C104). We selected these data sources for several reasons. First, they represent a standard set accessible to most public health authorities,[9,34] which may facilitate replication of the modeling approach elsewhere. Second, they contain features across several domains (e.g., physical environment, social capital) known to be associated with neighborhood-level overdose mortality rates.[8] And third, they were accessible at the CBG level to facilitate predictions that would guide CBG-level intervention.[8,9]

The model outcome was unintentional overdose deaths that occurred in RI between 1 January, 2016 and 30 June, 2020 obtained from RIDOH's State Unintentional Drug Overdose Reporting System (SUDORS).[24]

We used a variety of data sources as model predictors. First, EMS runs for nonfatal opioid overdoses from 1 January, 2016 to 30 June, 2020 were obtained from the RI EMS Information System.[35] Second, we obtained PDMP data from

1 July, 2016 to 30 June, 2020 from the RI PDMP, to capture counts of opioid analgesic and buprenorphine prescriptions dispensed and patients filling prescriptions at the CBG level. Using both prescription and patient data accounts for potential discordance between neighborhood of pharmacy and neighborhood of patient residence.[26]

Third, we extracted 5-year ACS estimates for calendar years 2016–2020 from the US Census.[27] Finally, we derived public access land use, health care, and social service data from a range of sources including the Substance Abuse and Mental Health Services Administration's treatment locator,[28] Brown University's PolicyMap license,[29] SimplyAnalytics,[30] the RI Department of Business Regulation,[31] RIDOH Licensing,[32] and the RI Geographic Information System, an open-source geospatial data hub.[33] For variables for which multiple years were available, we utilized the mean of a given variable's values across years.

Statistical Methods and Modeling Process

To predict future CBG-level overdose risk, we used two machine learning methods: gradient boosting machines and super learner. We selected these two methods through extensive testing of a range of approaches, detailed in eAppendix B; http://links.lww.com/EDE/C104. For clarity and brevity, we present only the results from our ensemble model and its composite base models (gradient boosting machines and super learner), each of which is also an ensemble model. We also compared all models' performance to a baseline model using the top ranked CBGs by number of overdose deaths during the training periods to predict future overdose deaths.

Gradient boosting machines are a tree-based ensemble method that offers an alternative to the more commonly used random forest algorithm.[36] Gradient boosting machines are useful for modeling complex relationships using high-dimensional data. Where random forest models construct an ensemble of deep, independent trees, gradient boosting machines build an ensemble of shallow trees with each subsequent tree building on the previous trees. We implemented gradient boosting machines using Python version 3.0 (Python Software Foundation, Wilmington, DE).

Super learner is an ensemble-based modeling approach that uses cross-validation to create a weighted optimal prediction from a library of a priori-specified candidate algorithms, where optimality is defined by the minimization of an objective function.[37] Our Super learner was a five-fold, cross-validated composite of elastic net, random forest, and gradient-boosted tree algorithms with an elastic net acting as a screening algorithm, and the objective function was minimizing mean squared error.[37] Super learner allows for integration of disparate modeling strategies to complement one another in solving a single prediction problem, with the ensemble super learner performing asymptotically as well as the best-performing candidate algorithm.[38] The library used in the

super learner model is presented in eAppendix C; http://links.lww.com/EDE/C104. We implemented super learner using R version 3.6.3 (R Foundation for Statistical Computing, Vienna, Austria).

Our final model was an ensemble of the gradient boosting machine and super learner base models, produced using a weighting approach. We calibrated ensemble weights to optimize our performance target, considering weights between 0 and 1 in increments of 0.02 to maximize the number of overdose deaths captured in the prioritized 20% of neighborhoods for each base model. We took this approach due to our highly specific model evaluation criteria selected a priori with RIDOH partners as part of the randomized trial within which this modeling study was embedded. Predictions were generated using the weighted linear combination of the base models that maximized this metric: 0.92 to gradient boosting machine and 0.08 to super learner. Schematics detailing training for these models are presented in eAppendices D and E; http://links.lww.com/EDE/C104.

Model Evaluation Criteria

We identified a priori modeling objectives and performance criteria with study partners at RIDOH and tailored these targets to the design and implementation of the randomized trial. Our modeling objective was to identify the highest-risk neighborhoods for overdose prevention resource allocation within municipalities, so we evaluated model performance using a custom evaluation metric developed in collaboration with RIDOH. Our development of the model evaluation criteria used for this study is detailed in a prior publication[39] and outlined below.

Our primary performance metric was defined a priori as the proportion of statewide overdose deaths that occurred in the 20% of statewide CBGs prioritized by the model, subject to the constraint below. RIDOH selected the ceiling of 20% to represent the percentage of neighborhoods that could reasonably be prioritized by overdose prevention organizations given existing resources. To facilitate randomization of municipalities into intervention and control conditions, we required that at least one neighborhood be prioritized by the model in each municipality, thus facilitating resource allocation decisions in every municipality assigned to the intervention condition. Given these constraints, we established the following benchmark model performance a priori through close collaboration with RIDOH: the model must meet or exceed a threshold of 40% of predicted statewide overdose deaths (contained in the prioritized 20% of CBGs) before use in the trial.

We also considered secondary performance metrics to assess equity in neighborhood-level resource allocation across several dimensions, including urbanicity, racial/ethnic segregation, and neighborhood poverty. Given patterns of segregation in RI, with urban CBGs more racially diverse than nonurban CBGs,[23] we considered racial and socioeconomic equity separately by urbanicity. Within urban jurisdictions, we considered the proportion of CBGs selected for resource allocation by neighborhood segregation level, identified using percent non-White and Theil's H as a multigroup entropy index.[40] We considered neighborhoods as majority White, majority non-White, or mixed. Within nonurban jurisdictions, we considered the proportion of CBGs selected for resource allocation by neighborhood poverty level. Consistent with US Census definitions, we classified CBGs with more than 20% of the population living below the federal poverty line as high-poverty neighborhoods.[41]

Baseline Comparison Condition

To assess the performance of our models in comparison with neighborhood allocation of harm reduction resources based on past overdose burden, we compared our model performance to a "practice as usual" baseline condition, which approximates RIDOH's standard resource distribution practices given the availability and completeness of SUDORS overdose mortality data in RI. We defined this baseline condition as the 20% of CBGs statewide with the highest historical overdose death burden during the respective training periods, subject to the same constraint that at least one CBG per municipality be included. We assessed the CBGs included in the baseline condition using the same model evaluation criteria described above.

Model Training and Testing

We used 6-month prediction windows, established in collaboration with RIDOH, as a realistic time period for RIDOH and community-based organizations to adjust harm reduction resource allocation. Our training period of 1 January, 2016–30 June, 2020 afforded us a total of seven 6-month windows. The spatiotemporal distribution of overdose deaths in RI varied between 6-month periods in the training data, owing to the rarity of the outcome. Due to this variation, we sought to increase generalizability and reduce bias due to overfitting by averaging performance across two test periods, rather than a single test period.[42] Therefore, we utilized the first five 6-month windows (1 January, 2016–30 June, 2019) as training periods and 1 July–31 December, 2019 and 1 January–30 June, 2020 as the two testing periods.

We implemented a moving window approach to construct an ensemble of gradient boosting machine and super learner base model predictions. We predicted each target period t using data from $t-1$ and $t-2$ to make predictions. Base models thus rolled forward across the available training periods until all data were exhausted. For prelaunch model training, this produced five base models and two held out test periods.

For the gradient boosting machine base model, we simultaneously utilized all available features corresponding to the respective $t-1$ and $t-2$ time periods. In order to facilitate predictions based on unique sources of variation, the super learner modeling approach relied on only data from the $t-1$ time period. To reduce dimensionality, we utilized elastic net feature selection screeners separately prior to fitting each

of the gradient boosting machine and super learner models.[43] Features selected for inclusion in super learner and gradient boosting machine base models are presented in eAppendices F and G; http://links.lww.com/EDE/C104.

Model Validation

To internally validate the model, we used SUDORS overdose mortality data from our trial launch period (1 July–31 December, 2020) and model update period (1 January–30 June, 2021). Our predictions for these periods were compared against the observed CBG-level overdose death counts after those data became available using our set of evaluation metrics. Model validation will remain ongoing across the life of the trial as future predictions are made every 6 months.

RESULTS

Overdose Mortality in Rhode Island

The overall rate of overdose mortality in RI increased across the study time period, from 29.4 per 100,000 residents in 2016 to 33.9 per 100,000 residents in 2020. The median CBG overdose death count across the full study period was 1, with interquartile range 0–2 and overall range 0–21.

Test Performance

Table 1 presents the test performance of the ensemble model (weighted average of gradient-boosted machine and super learner base models) across our primary metric, the proportion of statewide overdose deaths captured in the prioritized 20% of CBGs, constraining the model to select at least one CBG per municipality. The ensemble model pre-trial launch test average was 40.2% of overdose deaths across the periods of 1 July–31 December, 2019 and 1 January–30 June, 2020, as compared to 39.5% for the gradient boosting machine model, 34.1% for super learner, and 33.5% for the baseline condition.

Table 2 presents the test performance of the ensemble model and baseline comparison across our secondary metrics. Statewide, the ensemble model prioritized 23.3% of urban and 15.6% of nonurban CBGs. Within urban jurisdictions, it prioritized 27.2% of racially integrated, 54.4% of majority non-White, and 4.5% of majority White CBGs. Within nonurban jurisdictions, it prioritized 14.2% of low-poverty and 27.8% of high-poverty CBGs. Compared to the baseline condition, this reflects increased proportions of majority non-White and integrated urban CBGs and a decreased proportion of high-poverty nonurban CBGs. Secondary performance metrics for gradient boosting machine and super learner models are available in eTables 1 and 2; https://links.lww.com/EDE/C105.

Validation Performance

Table 3 presents the validation performance of the ensemble model and base models along our primary metric. Our ensemble model exceeded the a priori threshold and successfully prioritized 20% of CBGs that captured 44.1% of all overdose deaths in the subsequent 6-month period. This performance was comparable to the gradient boosting machine base model (44.3% of overdose deaths captured) and higher than super learner (33.1%) and the baseline condition (36.2%).

Table 4 presents predictive performance of the ensemble model across our secondary metrics. Statewide, the ensemble

TABLE 1. Model Test Performance: Proportion of Overdose Deaths Captured

Proportion of Overdose Deaths Captured at 20% of CBGs Prioritized

	Test Period 1 (1 July–31 December, 2019)	Test Period 2 (1 January–30 June, 2020)	Test Average
Ensemble	37.0%	43.3%	40.2%
Gradient boosting machine	35.6%	43.3%	39.5%
Super learner	33.2%	34.9%	34.1%
Baseline: practice as usual	34.9%	32.1%	33.5%

Sources: Brown University[29]; Rhode Island Department of Business Regulation[31]; Rhode Island Department of Health[26]; Rhode Island Department of Health[24]; Rhode Island Department of Health[32]; Rhode Island Emergency Medical Services Information System[35]; Rhode Island Geographic Information System[33]; Substance Abuse and Mental Health Services Administration[28]; United States Census.[27]

TABLE 2. Ensemble Model and Baseline Comparison Test Performance: Health Equity Considerations

Proportion of CBGs Prioritized by Urban Designation

	Test Period 1 (1 July–31 December, 2019)	Test Period 2 (1 January–30 June, 2020)	Test Average
Ensemble model			
Urban CBGs	23.6%	23.0%	23.3%
Integrated	27.8%	26.5%	27.2%
Majority non-White	53.9%	54.9%	54.4%
Majority White	5.0%	4.0%	4.5%
Nonurban CBGs	15.2%	16.0%	15.6%
Nonpoverty areas	13.7%	14.7%	14.2%
Poverty areas	27.8%	27.8%	27.8%
Baseline: practice as usual			
Urban CBGs	21.0%	22.3%	21.7%
Integrated	21.6%	23.5%	22.6%
Majority non-White	41.2%	39.2%	40.2%
Majority White	10.4%	12.9%	11.7%
Nonurban CBGs	18.7%	16.9%	17.8%
Nonpoverty areas	16.0%	13.7%	14.9%
Poverty areas	41.7%	44.4%	43.1%

Sources: Brown University[29]; Rhode Island Department of Business Regulation[31]; Rhode Island Department of Health[26]; Rhode Island Department of Health[24]; Rhode Island Department of Health[32]; Rhode Island Emergency Medical Services Information System[35]; Rhode Island Geographic Information System[33]; Substance Abuse and Mental Health Services Administration[28]; United States Census.[27]

TABLE 3. Model Validation Performance: Proportion of Overdose Deaths Captured

Proportion of Overdose Deaths Captured at 20% of CBGs Prioritized

	Validation Period 1 (1 July–31 December, 2020)	Validation Period 2 (1 January–30 June, 2021)	Validation Average
Ensemble	42.4%	45.7%	44.1%
Gradient boosting machine	42.4%	46.2%	44.3%
Super learner	33.2%	33.0%	33.1%
Baseline: practice as usual	36.4%	36.0%	36.2%

Sources: Brown University[29]; Rhode Island Department of Business Regulation[31]; Rhode Island Department of Health[26]; Rhode Island Department of Health[24]; Rhode Island Department of Health[32]; Rhode Island Emergency Medical Services Information System[35]; Rhode Island Geographic Information System[33]; Substance Abuse and Mental Health Services Administration[28]; United States Census.[27]

model prioritized 23.7% of urban and 15.1% of nonurban CBGs. Within urban jurisdictions, it prioritized 29.0% of racially integrated, 52.5% of majority non-White, and 5.0% of majority White CBGs. Within nonurban jurisdictions, it prioritized 13.1% of low-poverty and 32.0% of high-poverty CBGs. Compared to the baseline condition, this reflects increased proportions of majority non-White and integrated urban CBGs, and high-poverty nonurban CBGs. Secondary validation metrics for gradient boosting machine and super learner models are available in eTables 3 and 4; https://links.lww.com/EDE/C105.

DISCUSSION

We developed and internally validated an ensemble machine learning model to predict overdose risk at neighborhood level in RI. For both testing and validation, our model, an ensemble of gradient boosting machine and super learner base models, successfully predicted over 40% of statewide overdose deaths within the top 20% of CBGs, the benchmark for success determined a priori with RIDOH.[8] Our findings indicate that spatiotemporal forecasting of neighborhood-level overdose mortality is feasible as a strategy to inform overdose prevention resource allocation to a degree of accuracy suitable for practitioners.

Our models' performances varied somewhat as the available training data increased, with ensemble and gradient boosting machine models performing comparably at validation and super learner consistently underperforming the gradient boosting machine and ensemble. Gradient boosting machine models marginally outperformed the ensemble model during the second validation period, while the ensemble demonstrated a marginal gain in overdose capture at prelaunch testing. We selected the ensemble model for the PROVIDENT trial because it achieved the performance benchmark established with RIDOH in advance of the trial launch, while gradient boosting machine and super learner standalone models did not. This suggests that the ensemble model may offer greater potential when training periods are limited, while the gradient boosting machine's increasing performance across the validation periods is suggestive of its potential to guide public health interventions in jurisdictions with more extensive training data.[44] It is possible that the use of a single time period in super learner's moving windows, compared with two time periods for gradient boosting machine, may have contributed to its lower performance.

TABLE 4. Ensemble Model and Baseline Comparison Validation Performance: Health Equity Considerations

Proportion of CBGs Prioritized by Urban Designation

	Validation Period 1 (1 July–31 December, 2020)	Validation Period 2 (1 January–30 June, 2021)	Validation Average
Ensemble model			
Urban CBGs	23.4%	24.0%	23.7%
Integrated	29.6%	28.4%	29.0%
Majority non-White	50.0%	54.9%	52.5%
Majority White	5.0%	5.0%	5.0%
Nonurban CBGs	15.5%	14.6%	15.1%
Nonpoverty areas	13.4%	12.7%	13.1%
Poverty areas	33.3%	30.6%	32.0%
Baseline: practice as usual			
Urban CBGs	23.0%	21.2%	22.1%
Integrated	20.4%	19.8%	20.1%
Majority non-White	45.1%	40.2%	42.7%
Majority White	13.9%	12.9%	13.4%
Nonurban CBGs	16.0%	18.4%	17.2%
Nonpoverty areas	15.0%	16.3%	15.7%
Poverty areas	25.0%	36.1%	30.6%

Sources: Brown University[29]; Rhode Island Department of Business Regulation[31]; Rhode Island Department of Health[26]; Rhode Island Department of Health[24]; Rhode Island Department of Health[32]; Rhode Island Emergency Medical Services Information System[35]; Rhode Island Geographic Information System[33]; Substance Abuse and Mental Health Services Administration[28]; United States Census.[27]

Critically, all our predictive models outperformed the baseline condition during the prelaunch test periods, and all but super learner outperformed the baseline at validation, suggesting that the use of machine learning to inform intervention distribution may offer gains over public health practice as usual. We will continue to internally validate our ensemble model and the two base models across the life of the trial to assess differences in long-term stability of their predictions.

Crucially, this study contributes to emerging literature incorporating predictive analytics into population-level overdose prediction. Where prior studies have identified neighborhood-level predictors of overdose mortality[45,46] and assessed the suitability of machine learning to inform the distribution of public health interventions,[47] ours is the first to develop and validate a spatiotemporal machine learning model to predict neighborhood overdose risk for public health practice. By foregrounding prediction for public health prevention, our work builds on the robust and growing body of research utilizing predictive analytics to guide clinical practice and reduce patient-level overdose risk,[18,19,48] and introduces a possible tool for public health authorities to integrate into practice. This proactive approach to overdose prevention, in contrast to resource distribution based solely on retrospective area-level overdose history, presents a potential paradigm shift for public health practice. Likewise, our approach differs from methods using publicly sourced web data (e.g., Google Flu Trends[49]), which may be prone to error, by analyzing data sources that comprise "gold standard" overdose surveillance and are generally housed in state and local health departments. Moreover, using such data sources can facilitate uptake of area-level predictive modeling by practitioners. Demonstrating the potential for a proactive approach is essential as health authorities seek new strategies to address rising overdose deaths and spatially shifting patterns of risk.

LIMITATIONS

This study is subject to several limitations. First, while machine learning offers a methodologic toolkit for neighborhood-level forecasting, accuracy may be limited by this study's small training set. As predictions are spatiotemporal, the relatively low number of observation periods available for training may introduce bias if changes in the spatiotemporal distribution and risk predictors of overdose occur quickly at CBG level.[50] However, access to the full population of CBGs for model training may inform model accuracy in the absence of additional time periods, as demonstrated through our internal validation.

Second, this modeling study is embedded in a randomized trial to measure the effect of proactive, prediction-driven prevention resource allocation on fatal and nonfatal overdose. Thus, if the trial is successful in reducing overdose, outcome data will be affected in future time periods. However, given that this modeling study presents only the trial launch and early validation predictions, with increasing accuracy along

our a priori selected primary metric, this bias, if present, is unlikely to affect our short-run model predictions.

Third, data availability may introduce selection bias in predictions. Some data sources may only signal community-level overdose risk as captured by service-involved populations (e.g., buprenorphine treatment data from the PDMP). Likewise, use of ACS estimates, which vary little across time periods, may prioritize neighborhoods with endemically high rates of overdose associated with known risk factors (e.g., poverty), while not capturing short-run spatial variation in overdose. Data sources not accessible for use during modeling but with established signals for overdose risk, for example, methadone treatment data,[51] may prioritize out-of-treatment populations in model predictions. While emerging research demonstrates the capacity for social media data to inform area-level overdose risk,[52] we restricted our data sources to those widely accessible to public health practitioners.

Fourth, as a tool for public health practice, the wealth of data sources available for use in RI to build our model may limit its portability to other jurisdictions. Application of these approaches in other settings will be crucial to assess its utility as a public health tool.

Fifth, our model evaluation criterion identified only the top 20% of CBGs for public health prioritization, a determination made in concert with practitioners at RIDOH. We considered other approaches (e.g., prioritization based on rank order of CBGs within each municipality), but these were deemed impractical to implement by state health authority and community-based practitioners who were research partners. While the single threshold facilitated feasibility in implementation of resource targeting by practitioner partners, allocating resources based on a fixed threshold may be inefficient relative to future risk. Future work could explore the capacity for modeling across a continuous and dynamic risk threshold, with practitioner investments relative to the predicted future risk.

Sixth, since our model was embedded in a randomized trial, with municipality as the unit of randomization,[8] our evaluation required that at least one CBG be selected in each municipality in RI. In addition, the state health department required prioritization of at least one CBG per town in order to ensure that resources were not being directed toward a small number of high burden municipalities as a result of the trial. Thus, while this may introduce inefficiency in public health resource targeting from a statewide perspective, it was viewed as more equitable from a municipal and health department perspective. Prior research conducted by our study team has illustrated the relative tradeoffs between the targeting composition of neighborhoods within and across municipalities.[39]

Relatedly, to achieve the 40% performance threshold identified a priori by RIDOH for this trial, our model prioritized predictive performance over model transparency. Likewise, our ensembling approach was tailored to these unique model evaluation criteria and benchmarks for success in the context of a randomized trial. Future work not

subject to such constraints could consider the relative benefits of a more transparent modeling approach. Further, prior literature has demonstrated the successful implementation of super learner in constrained optimization settings.[53] Future work could directly integrate site-specific loss functions into modeling as these approaches expand in the applied public health sector.

Seventh, our second test period and both validation periods used data from the post-COVID era, while training data were drawn from the pre-COVID era. It is possible that our data and predictions may be subject to bias due to changes in the substance use and harm reduction services landscape during the SARS-CoV-2 pandemic. However, our use of two validation periods extending beyond the first wave of the pandemic imbues confidence in our model.

CONCLUSIONS

This study presents the development and internal validation of an ensemble machine learning model to predict neighborhood-level overdose risk in RI. Our ensemble model achieved the target performance during test and validation phases and outperformed a baseline condition representing standard public health practice. We are currently testing the effect of using the model to guide overdose prevention resource distribution through a randomized trial. As the overdose epidemic continues, area-based machine learning models have the potential to inform prevention proactively, offering a new paradigm for intervention in jurisdictions impacted by the overdose crisis. Future work should consider application of our ensemble modeling approach in jurisdictions with profiles that differ from RI, as well as the inclusion of additional data sources to inform population-level predictive modeling for overdose prevention practice.

ACKNOWLEDGMENTS

REFERENCES

1. Spencer MR, Miniño AM, Warner M. Drug overdose deaths in the United States, 2001-2021. *NCHS Data Brief*. 2022;457:1–8.
2. Mattson CL, Tanz LJ, Quinn K, Kariisa M, Patel P, Davis NL. Trends and geographic patterns in drug and synthetic opioid overdose deaths - United States, 2013-2019. *MMWR Morb Mortal Wkly Rep*. 2021;70:202–207.
3. Radfar SR, De Jong CAJ, Farhoudian A, et al; ISAM-PPIG Global Survey Consortium. Reorganization of substance use treatment and harm reduction services during the COVID-19 pandemic: a global survey. *Front Psychiatry*. 2021;12:639393.
4. Cartus AR, Li Y, Macmadu A, et al. Forecasted and observed drug overdose deaths in the US during the COVID-19 pandemic in 2020. *JAMA Netw Open*. 2022;5:e223418.
5. Haffajee RL, Sherry TB, Dubenitz JM, et al. *U.S. Department of Health and Human Services Overdose Prevention Strategy*. Office of the Assistant Secretary for Planning and Evaluation, US Department of Health and Human Services; 2021.
6. Saloner B, McGinty EE, Beletsky L, et al. A public health strategy for the opioid crisis. *Public Health Rep*. 2018;133:24S–34S.
7. Wilt GE, Lewis BE, Adams EE. A spatial exploration of changes in drug overdose mortality in the United States, 2000-2016. *Prev Chronic Dis*. 2019;16:E33.
8. Marshall BDL, Alexander-Scott N, Yedinak JL, et al. Preventing Overdose Using Information and Data from the Environment (PROVIDENT): protocol for a randomized, population-based, community intervention trial. *Addiction*. 2022;117:1152–1162.
9. Marshall BDL, Yedinak JL, Goyer J, Green TC, Koziol JA, Alexander-Scott N. Development of a statewide, publicly accessible drug overdose surveillance and information system. *Am J Public Health*. 2017;107:1760–1763.
10. Neill DB, Herlands W. Machine learning for drug overdose surveillance. *J Technol Hum Serv*. 2018;36:8–14.
11. Karim ME, Pang M, Platt RW. Can we train machine learning methods to outperform the high-dimensional propensity score algorithm? *Epidemiology*. 2018;29:191–198.
12. Olson RS, Cava W, Mustahsan Z, Varik A, Moore JH. Data-driven advice for applying machine learning to bioinformatics problems. *Pac Symp Biocomput*. 2018;23:192–203.
13. Marcus JL, Sewell WC, Balzer LB, Krakower DS. Artificial intelligence and machine learning for HIV prevention: emerging approaches to ending the epidemic. *Curr HIV/AIDS Rep*. 2020;17:171–179.
14. Balzer LB, Havlir DV, Kamya MR, et al. Machine learning to identify persons at high-risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. *Clin Infect Dis*. 2020;71:2326–2333.
15. Bharat C, Hickman M, Barbieri S, Degenhardt L. Big data and predictive modelling for the opioid crisis: existing research and future potential. *Lancet Digit Health*. 2021;3:e397–e407.
16. Ward PJ, Rock PJ, Slavova S, Young AM, Bunn TL, Kavuluru R. Enhancing timeliness of drug overdose mortality surveillance: a machine learning approach. *PLoS One*. 2019;14:e0223318.
17. Lo-Ciganic WH, Donohue JM, Hulsey EG, et al. Integrating human services and criminal justice data with claims data to predict risk of opioid overdose among Medicaid beneficiaries: a machine-learning approach. *PLoS One*. 2021;16:e0248360.
18. Lo-Ciganic WH, Huang JL, Zhang HH, et al. Evaluation of machine-learning algorithms for predicting opioid overdose risk among Medicare beneficiaries with opioid prescriptions. *JAMA Netw Open*. 2019;2:e190968.
19. Lo-Ciganic WH, 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–e465.
20. Dong X, Deng J, Hou W, et al. Predicting opioid overdose risk of patients with opioid prescriptions using electronic health records based on temporal deep learning. *J Biomed Inform*. 2021;116:103725.
21. Gellad WF, Yang Q, Adamson KM, et al. Development and validation of an overdose risk prediction tool using prescription drug monitoring program data. *Drug Alcohol Depend*. 2023;246:109856.
22. Diez Roux AV, Merkin SS, Arnett D, et al. Neighborhood of residence and incidence of coronary heart disease. *N Engl J Med*. 2001;345:99–106.
23. Rhode Island Statewide Planning Program. *Population counts and estimates*. Updated 2021. Available at: http://www.planning.ri.gov/planning-areas/demographics/data/data-downloads.php. Accessed March 28, 2022
24. Jiang Y, McDonald JV, Goldschmidt A, et al. State unintentional drug overdose reporting surveillance: opioid overdose deaths and characteristics in Rhode Island. *R I Med J (2013)*. 2018;101:25–30.
25. Rhode Island Department of Health. *Enhanced State Opioid Overdose Surveillance (ESOOS): CASE DEFINITION for Emergency Medical Services (EMS)*. Updated 2019. Available at: https://health.ri.gov/publications/guidelines/ESOOSCaseDefinitionForEMS.pdf. Accessed August 25, 2021
26. Rhode Island Department of Health. *Prescription Drug Monitoring Program*. Updated 2022. Available at: https://health.ri.gov/healthcare/medicine/about/prescriptiondrugmonitoringprogram/. Accessed December 17, 2021
27. US Census Bureau. *American Community Survey Data*. Updated 2021. Available at: https://www.census.gov/programs-surveys/acs. Accessed March 28, 2022

28. Substance Abuse and Mental Health Services Administration. *Treatment Locator Map: Providence, RI*. Updated nd. Available at: https://findtreatment.samhsa.gov/locator?sAddr=Providence,%20RI%2002908,%20 USA. Accessed April 1, 2020

29. Brown University. *PolicyMap*. Updated 2021. Available at: https://www.policymap.com/. Accessed January 15, 2020

30. Brown University. *SimplyAnalytics*. Updated 2021. Available at: http://app.simplyanalytics.com. Accessed December 4, 2020

31. Rhode Island Department of Business Regulation. *Commercial Licensing*. Updated 2022. Available at: https://dbr.ri.gov/commercial-licensing. Accessed January 13, 2021

32. Rhode Island Department of Health. *What We License*. Updated nd. Available at: https://health.ri.gov/licenses. Accessed January 13, 2021

33. Rhode Island Geographic Information System. *Rhode Island Geographic Information System (RIGIS)*. Updated nd. Available at: https://www.rigis.org/. Accessed July 16, 2018

34. Goldstick J, Ballesteros A, Flannagan C, Roche J, Schmidt C, Cunningham RM. Michigan system for opioid overdose surveillance. *Inj Prev*. 2021;27:500–505.

35. Hallowell BD, Chambers LC, Rhodes J, Basta M, Viner-Brown S, Lasher L. Using emergency medical services data to monitor nonfatal opioid overdoses in real time: development, validation, and use of a case definition, Rhode Island, 2018. *Public Health Rep*. 2021;136:40S–46S.

36. Zhang Z, Zhao Y, Canes A, Steinberg D, Lyashevska O; written on behalf of AME Big-Data Clinical Trial Collaborative Group. Predictive analytics with gradient boosting in clinical medicine. *Ann Transl Med*. 2019;7:152.

37. Naimi AI, Balzer LB. Stacked generalization: an introduction to super learning. *Eur J Epidemiol*. 2018;33:459–464.

38. van der Laan MJ, Polley EC, Hubbard AE. Super learner. *Stat Appl Genet Mol Biol*. 2007;6:Article25.

39. Allen B, Neill DB, Schell RC, et al. Translating predictive analytics for public health practice: a case study of overdose prevention in Rhode Island. *Am J Epidemiol*. 2023;192:1659–1668

40. Theil H, Finizza AJ. A note on the measurement of racial integration of schools by means of informational concepts. *J Math Sociol*. 1971;1:187–193.

41. US Census Bureau. *Poverty Glossary*. Updated 2021. Available at: https://www.census.gov/topics/income-poverty/poverty/about/glossary.html. Accessed March 28, 2022

42. Demšar J, Zupan B. Hands-on training about overfitting. *PLoS Comput Biol*. 2021;17:e1008671.

43. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. 2005;67:301–320.

44. Khoury MJ, Engelgau M, Chambers DA, Mensah GA. Beyond public health genomics: can big data and predictive analytics deliver precision public health? *Public Health Genomics*. 2018;21:244–250.

45. Bozorgi P, Porter DE, Eberth JM, Eidson JP, Karami A. The leading neighborhood-level predictors of drug overdose: a mixed machine learning and spatial approach. *Drug Alcohol Depend*. 2021;229:109143.

46. Schell RC, Allen B, Goedel WC, et al. Identifying predictors of opioid overdose death at a neighborhood level with machine learning. *Am J Epidemiol*. 2022;191:526–533.

47. Yedinak JL, Li Y, Krieger MS, et al. Machine learning takes a village: assessing neighbourhood-level vulnerability for an overdose and infectious disease outbreak. *Int J Drug Policy*. 2021;96:103395.

48. Guo J, Lo-Ciganic WH, Yang Q, et al. Predicting mortality risk after a hospital or emergency department visit for nonfatal opioid overdose. *J Gen Intern Med*. 2021;36:908–915.

49. Lazer D, Kennedy R, King G, Vespignani A. Big data the parable of Google Flu: traps in big data analysis. *Science*. 2014;343:1203–1205.

50. DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc*. 2020;27:2020–2023.

51. Sordo L, Barrio G, Bravo MJ, et al. Mortality risk during and after opioid substitution treatment: systematic review and meta-analysis of cohort studies. *BMJ*. 2017;357:j1550.

52. Matero M, Giorgi S, Curtis B, Ungar LH, Schwartz HA. Opioid death projections with AI-based forecasts using social media language. *NPJ Digit Med*. 2023;6:35.

53. Zheng W, Balzer L, van der Laan M, Petersen M; SEARCH Collaboration. Constrained binary classification using ensemble learning: an application to cost-efficient targeted PrEP strategies. *Stat Med*. 2018;37:261–279.