Title

Outcome class imbalance and rare events: An underappreciated complication for overdose risk prediction modeling.

Permalink
https://escholarship.org/uc/item/7q10c6f8

Journal
Addiction, 118(6)

Authors
Cerdá, Magdalena
Marshall, Brandon
Cartus, Abigail
et al.

Publication Date
2023-06-01

DOI
10.1111/add.16133

Peer reviewed

Published in final edited form as:
Addiction. 2023 June ; 118(6): 1167–1176. doi:10.1111/add.16133.

Outcome class imbalance and rare events: An underappreciated complication for overdose risk prediction modeling

Abigail R. Cartus, MPH PhD[1], **Elizabeth A. Samuels, MD MPH MHS**[1,2], **Magdalena Cerdá, DrPH**[3], **Brandon D.L. Marshall, PhD**[1,*]

[1]Department of Epidemiology, Brown University School of Public Health, Providence, Rhode Island

[2]Department of Emergency Medicine, Alpert Medical School of Brown University, Providence, Rhode Island

[3]Division of Epidemiology, Department of Population Health, Center for Opioid Epidemiology and Policy, School of Medicine, New York University, New York

Abstract

Background and aims—Low outcome prevalence, often observed with opioid-related outcomes, poses an underappreciated challenge to accurate predictive modeling. Outcome class imbalance, where non-events (*i.e.*, negative class observations) outnumber events (*i.e.*, positive class observations) by a moderate to extreme degree, can distort measures of predictive accuracy in misleading ways and make the overall predictive accuracy and the discriminatory ability of a predictive model appear spuriously high. We conducted a simulation study to measure the impact of outcome class imbalance on predictive performance of a simple SuperLearner ensemble model and suggest strategies for reducing that impact.

Design, Setting, Participants—Using a Monte Carlo design with 250 repetitions, we trained and evaluated these models on four simulated data sets with 100,000 observations each: one with perfect balance between events and non-events, and three where non-events outnumbered events by an approximate factor of 10:1, 100:1, and 1000:1, respectively.

Measurements—We evaluated the performance of these models using a comprehensive suite of measures, including measures that are more appropriate for imbalanced data.

Findings—Increasing imbalance tended to spuriously improve overall accuracy (using a high threshold to classify events vs. non-events, overall accuracy improved from 0.45 with perfect balance to 0.99 with the most severe outcome class imbalance), but diminished predictive performance was evident using other metrics (corresponding positive predictive value decreased from 0.99 to 0.14).

Conclusion—Increasing reliance on algorithmic risk scores in consequential decision-making processes raises critical fairness and ethical concerns. This paper provides broad guidance for

[*]Correspondence: Department of Epidemiology, Brown University School of Public Health, 121 S. Main Street, Providence, RI 02903, Brandon_marshall@brown.edu.

Declaration of interests: none

analytic strategies that clinical investigators can use to remedy the impacts of outcome class imbalance on risk prediction tools.

Keywords

Opioid; overdose; risk prediction; machine learning; predictive modeling

Introduction

The overdose crisis in the United States has intensified during the COVID-19 pandemic owing to a variety of factors including pandemic-related disruptions, social isolation, and an increasingly toxic supply of illicit drugs.(1, 2) Overdose mortality was steadily increasing through 2019, with a record 92,000 deaths in 2020(3) and 53,000 in the first six months of 2021 alone.(4) Algorithmic modeling to predict or identify patients at high risk of overdose and other adverse drug-related (particularly opioid-related) outcomes has been increasingly used to direct overdose prevention interventions. Many investigators have used both conventional approaches (typically logistic regression) and more novel machine learning methods to predict risk of opioid-related harms, such as fatal and nonfatal overdose, opioid-related hospitalizations, incident opioid use disorder, and persistent opioid use using large datasets including those from prescription drug monitoring programs,(5–9) commercial claims and electronic health records,(10–17) Medicare,(18, 19) Medicaid,(20) the Veterans Administration,(21, 22) and other administrative claims.(23)

Despite this growing literature, there is an underappreciated (yet addressable) problem of “outcome class imbalance”(24–26) which is associated with risk prediction modeling of rare events. Outcome class imbalance is a function of outcome prevalence and occurs in scenarios where non-cases (members of the so-called “negative” or “majority” outcome class) outnumber cases (members of the positive or minority outcome class) by a moderate to extreme degree and can produce low predictive model accuracy. Outcome class imbalance thus corresponds to the familiar situation of low outcome prevalence. This is an issue for any type of rare event, including opioid-related adverse events, which tend to be quite rare, especially in the large data sets that make attractive candidates for risk prediction modeling. In one study in Washington, for example, even in a relatively high-risk population of Medicaid enrollees receiving an opioid prescription, the cumulative incidence of opioid-related poisoning over five years was less than 0.5%.(27) Outcome class imbalance can distort some measures of predictive accuracy, in particular the *overall accuracy* measure, which is the total number of correct predictions or classifications made by the model divided by the total number of observations. If non-cases (negative or majority class) outnumber cases (positive or minority class) by a large or extreme degree, predictive models, whether regression or more complex machine learning models, can achieve excellent overall predictive accuracy by simply classifying most or all observations as non-events. For example, in a data set of 100 observations with 99 non-events and 1 outcome event, a risk prediction model could achieve an overall accuracy of 0.99 by classifying every observation as a non-event. Among other solutions we will discuss later, purposeful sampling may be employed to reduce the degree of imbalance.

How much outcome class imbalance affects predictive accuracy depends on a number of factors, including the complexity and noisiness of the data used to develop the model.(28) However, so-called "singular" or threshold-dependent assessment metrics of predictive performance, especially those that incorporate information about the distribution of positive and negative outcome classes relative to one another (like overall accuracy), may generally be expected to yield misleading results in imbalanced data sets. Threshold-free, curve-based metrics generally give a more comprehensive picture of predictive performance than a single threshold-specific accuracy measure. For example, receiver operating characteristics (ROC) curves visualize the tradeoff between true positives and false positives (sensitivity and 1-specificity) across the range of all possible cutoffs and thus give a more global picture of algorithm performance. However, ROC curves are also sensitive to outcome class distribution and may thus also mislead.(29) Therefore, designing risk prediction algorithms for opioid-related outcomes represents a scenario where imbalance is likely (at a minimum, not accounting for other data complexity issues) and where performance is most commonly evaluated using overall accuracy and ROC curve analysis.

Risk prediction for opioid-related harms is a growing area of substance use epidemiology, and risk scores are being used in clinical care in, as just one example, the form of the NarxCare score, an algorithmic risk score built into prescription drug monitoring program (PDMP) software interfaces offered by Appriss, Inc.(30) While the inputs, development, and performance of the NarxCare scores are proprietary, available reports indicate that their high discriminatory (ROC) performance is used as evidence of these scores' accuracy and effectiveness(31) – evidence that has been called into question by more detailed analyses of the scores' performance.(32) This performance accuracy may also be misleading if the NarxCare scores are tested, generated, or used in highly imbalanced data, such as the data from PDMP databases. We demonstrate below that the literature on risk prediction for opioid-related harms may paint a misleading picture as to the accuracy and clinical utility of these algorithms.

Here, we conducted a simulation study to illustrate the effects of outcome class imbalance on a wide range of performance metrics for two simple prediction models (logistic regression and random forest). We did not seek to generate "accurate" risk predictions. Rather, our objective was to illustrate the effects and potential pitfalls of outcome class imbalance in predictive modeling studies and to offer investigators some considerations for building, interpreting, and reporting on risk prediction models for opioid-related adverse events and other rare outcomes in clinical medicine and public health.

Methods

To inform design of the simulated data sets, we first reviewed outcome event rates in a convenience sample of published papers predicting opioid-related adverse effects. We selected 20 recent papers reporting on risk prediction models across a range of opioid-related outcomes and using a variety of data sources (Table 1).(5–23, 33) Because our initial search strategy yielded relatively few papers, we selected papers reviewed in a Tseregounis *et al*. review of risk prediction analyses for opioid research and a selection of other papers identified from a PubMed search of "risk prediction" AND "opioid."(34) We compiled information including year of publication, prediction target/outcome, the data source used, the cumulative incidence of the outcome, prediction window, analytic approach, and the order of magnitude of the outcome class imbalance expressed as a ratio (*e.g.*, 10:1 for a cumulative incidence of the outcome indicating that non-cases outnumbered cases in the data by an approximate factor of 10).

Next, we used a Monte Carlo design to simulate 250 replications of four data sets with 100,000 observations each. For each degree of outcome class imbalance (1:1, 10:1, 100:1, 1000:1), we thus simulated 250 datasets with N = 100,000. We simulated the data from a logistic regression model to approximate the outcome of fatal overdose (in which it is not possible for the same subject to have multiple outcome events). The 1:1 data set had perfect balance (cases and non-cases accounted for 50% of the data each) and was included to serve as a reference to compare against model performance in the more imbalanced data sets. Furthermore, in each data set, we simulated 20 covariates (10 binary and 10 continuous covariates) with coefficients of a similar magnitude to what can be found in the opioid-related risk prediction literature (coefficients ranging from 0.70 to 3.5).(34) Visual inspection of the probability curves of the continuous covariates indicated a broadly linear trend and as such, more complex functional forms (*e.g.*, splines or polynomials) were not considered.

Each replication of each data set (in each of the four categories of outcome class imbalance) were prepared for modeling by first being split into training (N = 80,000) and test (N = 20,00) sets (using package *caret*(35)) while ensuring a similar outcome prevalence in both sets. We then trained a SuperLearner ensemble algorithm with ten-fold cross-validation on each training set. We used SuperLearner to “stack” three basic algorithms: the simple mean, a generalized linear model, and penalized linear regression. It is possible that the penalized regression base learner may have chosen different covariate mixes in different simulation runs. Other base learners for SuperLearner are available; due to limited computing resources and time and due to the illustrative nature of this analysis we chose very simple models, but others can be explored for other applications.

We used the parameters from each model built on the training set to generate predictions from the corresponding test set. From each model, we thus generated a vector of predicted risk scores (continuous probabilities between 0 and 1). We visualized the distribution of risk scores for each model as frequency histograms. We also generated both receiver operating characteristic (ROC) and precision-recall curves (similar to ROC curves, but showing the positive predictive value or precision versus the sensitivity or recall across all possible thresholds) for each model across each degree of outcome class balance.

In order to generate measures of predictive accuracy based on the predicted and true binary classifications arranged into a 2×2 table (also called a *confusion matrix*), it was necessary to choose thresholds of the risk score distribution to classify the predicted risk scores as cases or non-cases. We chose three thresholds for the primary analysis: the 50th, 75th, and the 99th percentile of each risk score distribution. Any observations with a predicted risk score greater than the 50th, 75th, or 99th percentile of the risk score distribution was classified as a case, while the rest were classified as non-cases. Using these predicted classifications, we constructed confusion matrices (2×2 tables of predicted vs. true binary classifications) to derive several measures of predictive accuracy. The measures we calculated were sensitivity (probability that a true case is classified as a predicted case), specificity (probability that a true non-case is classified as a predicted non-case), positive predictive value (probability that a predicted case is a true case), negative predictive value (probability that a predicted non-case is a true non-case), and overall accuracy (the total number of correct classifications divided by the total number of observations).We also used these classifications to calculate Brier scores to assess model calibration. The Brier score is the mean squared difference between the actual outcome of a single observation and the predicted probability of the outcome assigned to that observation.(36) Higher Brier scores thus indicate poorer calibration and lower Brier scores indicate better calibration.

We also visualized the confusion matrices (2×2 tables of true and predicted classifications) from one replication to clarify the distribution of false positives and false negatives and illustrate how changing the classification threshold affects their distribution.

We referenced extensions to the CONSORT and STROBE guidelines for simulation studies to guide the presentation of the methods and results included here.(37)

Results

We first reviewed outcome prevalence in 20 recent papers that developed risk prediction models for opioid-related harms (*e.g.*, fatal opioid overdose, Table 1).(5–23, 33) In these papers, non-cases outnumbered cases by a factor of at least 100 unless a deliberate strategy was used to over-sample outcome events; of the 20 papers we chose, 4 (20%) used sampling which reduced the degree of outcome class imbalance. The most common data sources for risk model development were administrative/claims data (10 papers) and prescription drug monitoring program data (6 papers) and most of the prediction windows were 1–2 years.

Each simulated data set had 100,000 observations but a different number of “cases” approximately corresponding to outcome class balance ratios of 1:1, 10:1, 100:1, and 1000:1, respectively. Histograms of the risk scores generated by each model illustrated the impact of outcome class imbalance (Figure 1). When the ratio of non-cases to cases was 1:1, the risk scores were distributed approximately evenly between 0 and 1; as the imbalance ratio increases, the risk scores become increasingly clustered close to 0, with a long right tail. This suggests that risk scores tend to be closer to zero when imbalance is more severe. These histograms represent results from just one of the 250 simulation runs (the 100th); some variability in the exact frequencies across simulation runs is to be expected, but the general trend is the same.

As the degree of outcome class imbalance increased, ROC and precision-recall performance changed in opposite directions (Figure 1). In general, the ROC curves exhibited consistent good performance as outcome class imbalance became more severe, though there was more variability across simulation runs at higher levels of imbalance. However, area under the precision-recall curve was markedly reduced even with the lowest degree of outcome class imbalance (10:1) and progressively worsened as the degree of imbalance increased.

When using risk score cutoffs to classify each observation in the test set as a case or non-case, Brier scores assessing model calibration decreased as the degree of imbalance increased, indicating that calibration improved, likely spuriously, with more severe outcome class imbalance. Changing degrees of outcome class imbalance and changing risk score cutoffs yielded patterns in the performance metrics. Across each risk score cutoff (50th, 75th, and 99th percentiles of the risk score distribution), increasing outcome class imbalance corresponded to increased sensitivity and negative predictive value and decreased specificity and positive predictive value (except at the 99th percentile threshold, where specificity is constant and high across all levels of outcome class imbalance). At the 50th percentile threshold, sensitivity and negative predictive value were higher overall (across all degrees of outcome class imbalance) and positive predictive value and specificity were lower overall. At the 50th percentile risk score cutoff, overall accuracy decreased with increasing outcome class imbalance; at the 75th percentile risk score cutoff, overall accuracy did not change in a consistent direction, and at the 99th percentiles cutoff, overall accuracy actually increased with more severe outcome class imbalance (Table 1).

Confusion matrices from one simulation run show the distribution of true and false positive and negative classifications across degrees of outcome class imbalance and risk score thresholds (Table 2). Across all risk score thresholds, increasing outcome class imbalance resulted in a decreased number of true positive and false negative classifications and an increased number of false positive and true negative classifications. The absolute number of classifications in each quadrant varied according to the risk score threshold, regardless of degree of imbalance; because of the lower threshold, many more false positive classifications were observed at the 50th percentile cutoff while many more true negative classifications were observed at the 99th percentile cutoff. For example, as a percentage of the training set of 20,000 observations with perfectly balanced data, the proportion of false positives at the 50th, 75th, and 99th percentile classification thresholds was 0.75%, 0.02%, and 0%, respectively, while the number of false negatives was 14.75%, 39.01%, and 63.00%, respectively. When non-cases outnumbered cases by a factor of 10:1, the proportion of false positives at the same increasing thresholds was 18.71%, 1.92%, and 0.01%, while the corresponding proportion of false negatives was 0.33%, 8.55%, and 30.64%. With the most severe outcome class imbalance (1000:1), the proportion of false positives was 46.79%, 21.86%, and 0.42% and the proportion of false negatives was 0.005%, 0.08%, and 2.64% across the same increasing classification thresholds. As with the risk score histograms in Figure 1, these confusion matrices represent results from one simulation run (the 100th); exact proportions will vary across runs but the general trend is consistent.

In a setting where a risk prediction algorithm is used to guide clinical decision making, this could result in a situation where the high apparent predictive accuracy of a risk scoring algorithm (*e.g.*, for fatal overdose) represents the increasing number of true negatives correctly classified by the algorithm. This could pose a problem especially where high risk score cutoffs are used, for example, to identify patients in the 99th percentile or higher of predicted risk scores. The high overall accuracy observed in this scenario could correspond to the algorithm’s excellent performance at predicting true negatives and a relatively small number of false positives; while the performance would technically be accurate, this would not be informative as to an individual patient's risk of adverse outcomes based on their risk score.

Discussion

The results of our simulation study illustrated the impact of progressively more pronounced outcome class imbalance on the performance of two predictive models. As expected, the impacts of outcome class imbalance were particularly evident in decreased positive predictive value and reduced areas under the precision-recall curves, and paradoxically also evident in increased overall accuracy, area under the ROC curve, and calibration. Overall accuracy in particular is sensitive to more than just the class distribution. Though in general, overall accuracy will improve with increasing outcome class imbalance, this may not be the case if a lower classification threshold is chosen—this will result in false positives (incorrect positive classifications) that drive the overall accuracy back down. We summarize recommendations for researchers conducting risk prediction analyses and for readers tasked with interpreting the results of these studies below.

First, we have demonstrated that threshold-free curve analyses are generally more informative than singular measures; of these, our results confirm that precision-recall plots are more informative than ROC curves in the presence of imbalanced data.(38) In general, we observed consistently high area under the ROC curve as the degree of outcome class imbalance increased and even greatly improved overall accuracy (which corresponds to area under the ROC curve) in a high-imbalance context when a higher risk score threshold for classification is used, which can be highly misleading. By contrast, the precision-recall curves registered the impact of outcome class imbalance even at the lowest degree of imbalance (10:1). This is because ROC curves are not sensitive to class distribution, whereas the precision-recall curve shows the relationship between positive predictive value and sensitivity across all possible thresholds and thus incorporates information about the outcome prevalence (the area under the precision-recall curve depends on the baseline outcome prevalence).(38) We also demonstrated that choice of classification threshold can affect the performance of a predictive model and change how the impact of outcome class imbalance is registered in the predictive performance. With a lower (in our case, less appropriate) classification threshold, each model generated more false positives, which translated into less implausibly high overall accuracy values but substantially damaged the positive predictive value. Unlike singular assessment metrics, curve-based analyses do not depend on setting a classification threshold and can be used to identify an optimal threshold (*e.g.*, one that maximizes both sensitivity and 1-specificity, or both sensitivity and positive predictive value). In sum, we recommend that risk prediction studies report the results of threshold-free curve analyses, and further recommend precision-recall curves in addition to or instead of ROC curves in the context of imbalanced data.

Second, researchers should consider analytic strategies to handle outcome class imbalance. While the most appropriate analytic strategy depends on the research question, objectives, and performance measures of most interest, sampling and cost-sensitive learning approaches represent two possible analytic approaches to handling outcome class imbalance. A fuller explication of analytic approaches to outcome class imbalance for predictive modeling may be found elsewhere(28, 39); here, we present some cursory considerations and definitions to aid researchers who are interested in exploring further.

Sampling approaches to outcome class imbalance vary in complexity but share the same fundamental idea: sampling or resampling an imbalanced data set to achieve more balance. At the simplest end are “naïve” sampling approaches: random undersampling (discarding non-outcome or majority class observations to achieve balance) and random oversampling (duplicating outcome or minority class events to achieve balance). These simple approaches do have drawbacks; undersampling discards information and may not be feasible (for example, in a data set with few outcome events), while oversampling duplicates existing observations and can thus lead to overfitting. On the other hand, these techniques are accessible and easy to implement; they may be combined to achieve a satisfactorily balanced data set, and many extensions and refinements are available.(40–46) Importantly, because sampling distorts the marginal outcome prevalence, investigators who choose a sampling approach will need to calculate sampling weights and apply them to computation of performance measures. If weighting is not feasible, sensitivity and specificity of an algorithm built on sampled data should not be reported, as these metrics depend on the marginal outcome prevalence and lose their meaning if the analytic data have been sampled to achieve a particular outcome prevalence.

In contrast to sampling approaches, which alter the data set, cost-sensitive learning approaches(47) allow investigators to impose higher “costs” for misclassifying certain (*e.g.*, positive class) observations. By default, predictive algorithms assume that false-positive and false-negative classifications have equal costs. However, this may not be true, and the cost of misclassifying a patient as high- or low-risk for future overdose or another adverse outcome has clinical (in addition to mathematical) relevance. In opioid-related risk prediction modeling, the clinical costs of misclassification are dynamic and context-specific; a false negative may be more clinically costly if it leads to opioid prescribing which may increase overdose risk, either in dosage, duration, or medication combinations. A false positive may be more clinically costly if the “high risk” designation creates a barrier to accessing needed medications for individuals with complex medical conditions. Cost-sensitive learning approaches can incorporate these differential costs into the model to, for example, penalize false positives more than false negatives, according to the objectives of the research.

Finally, whatever analytic approach is taken, we recommend examining and reporting a comprehensive suite of performance metrics. Precision-recall plots are not difficult to generate with available statistical software, and investigators may choose to present the F1 score (the harmonic mean of positive predictive value and sensitivity) in addition to or instead of overall accuracy. It is important to evaluate the calibration of a predictive model (how well the predicted outcomes accord with the observed outcomes) in addition to the discrimination (how well the model is able to distinguish cases from non-cases). (29, 34) However, it is noteworthy that per our results, the Brier score evaluating model calibration improves with increasing outcome class imbalance, suggesting that calibration results may also be misleading in the context of imbalanced data. To ensure a comprehensive understanding of the performance of a given predictive model, we recommend evaluating the metrics reported in Table 2 (at minimum) and at least one curve-based metric (ROC, precision-recall, or cost curves).

In summary, increasing reliance on algorithmic risk scores in consequential decision-making processes in medicine and other fields (e.g., pretrial detention, child protective services investigations, and more) raises critical fairness and ethical concerns.(48, 49) Algorithmic patient-level risk scores for adverse outcomes like opioid overdose or misuse are integrated into many prescription drug monitoring program software products (PDMPs are now operating in every US state except Missouri).(50) These risk scores are proprietary (and opaque) but may nevertheless be used in consequential prescribing decisions. For example, opioid dose tapering (such as might be an intuitive approach for a patient with a high risk score)(32, 50) is associated with increased risk of overdose.(51) Outcome class imbalance adds another layer of analytic complexity to the already challenging task of predictive modeling for opioid-related risks, and investigators should be alert to the presence and impact of outcome class imbalance in their data to avoid building a predictive model with misleading, inaccurate performance. Future research in this area should be particularly attentive to the potential for differential impacts of outcome class imbalance across subgroups (*e.g.*, racial/ethnic or socioeconomic groups). With the high-stakes consequences and real impact on people's lives that algorithmic risk scores can have, researchers have a responsibility to build and evaluate accurate predictive models in a comprehensive and transparent way.

Acknowledgements:

We would like to thank Dr. William Goedel for his extensive and indispensable assistance with running the simulations for the study on the available supercomputer.

Sources of Funding:

ARC, MC, and BDLM are supported by the National Institute on Drug Abuse (R01-DA046620).

References

1. Macmadu A, Batthala S, Gabel AMC, Rosenberg M, Ganguly R, Yedinak JL, et al. Comparison of characteristics of deaths from drug overdose before vs during the COVID-19 pandemic in Rhode island. JAMA network open. 2021;4(9):e2125538–e. [PubMed: 34533569]
2. Kuehn BM. Accelerated overdose deaths linked with COVID-19. JAMA. 2021;325(6):523-.
3. Hedegaard H, Miniño A, Spencer M, Warner M. Drug overdose deaths in the United States, 1999–2020. NCHS Data Brief, No. 428. National Center for Health Statistics. 2021.
4. Control CfD Prevention. Drug overdose deaths in the US Top 100,000 annually. Atlanta: Centers for Disease Control and Prevention. 2021.
5. Ferris LM, Saloner B, Krawczyk N, Schneider KE, Jarman MP, Jackson K, et al. Predicting Opioid Overdose Deaths Using Prescription Drug Monitoring Program Data. Am J Prev Med. 2019;57(6):e211–e7. [PubMed: 31753274]
6. Ferris LM, Saloner B, Jackson K, Lyons BC, Murthy V, Kharrazi H, et al. Performance of a Predictive Model versus Prescription-Based Thresholds in Identifying Patients at Risk of Fatal Opioid Overdose. Subst Use Misuse. 2021;56(3):396–403. [PubMed: 33446000]
7. Chang HY, Krawczyk N, Schneider KE, Ferris L, Eisenberg M, Richards TM, et al. A predictive risk model for nonfatal opioid overdose in a statewide population of buprenorphine patients. Drug Alcohol Depend. 2019;201:127–33. [PubMed: 31207453]

8. Chang HY, Ferris L, Eisenberg M, Krawczyk N, Schneider KE, Lemke K, et al. The Impact of Various Risk Assessment Time Frames on the Performance of Opioid Overdose Forecasting Models. Med Care. 2020;58(11):1013–21. [PubMed: 32925472]

9. Saloner B, Chang HY, Krawczyk N, Ferris L, Eisenberg M, Richards T, et al. Predictive Modeling of Opioid Overdose Using Linked Statewide Medical and Criminal Justice Data. JAMA Psychiatry. 2020;77(11):1155–62. [PubMed: 32579159]

10. 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;33(10):1646–53. [PubMed: 29380216]

11. Ellis RJ, Wang Z, Genes N, Ma'ayan A. Predicting opioid dependence from electronic health records with machine learning. BioData Min. 2019;12:3. [PubMed: 30728857]

12. Dong X, Deng J, Hou W, Rashidian S, Rosenthal RN, Saltz M, 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. [PubMed: 33711546]

13. 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. [PubMed: 28340046]

14. Sun JW, Franklin JM, Rough K, Desai RJ, Hernandez-Diaz S, Huybrechts KF, et al. Predicting overdose among individuals prescribed opioids using routinely collected healthcare utilization data. PLoS One. 2020;15(10):e0241083. [PubMed: 33079968]

15. Cochran BN, Flentje A, Heck NC, Van Den Bos J, Perlman D, Torres J, et al. Factors predicting development of opioid use disorders among individuals who receive an initial opioid prescription: mathematical modeling using a database of commercially-insured individuals. Drug Alcohol Depend. 2014;138:202–8. [PubMed: 24679839]

16. Dong X, Rashidian S, Wang Y, Hajagos J, Zhao X, Rosenthal RN, et al. Machine Learning Based Opioid Overdose Prediction Using Electronic Health Records. AMIA Annu Symp Proc. 2019;2019:389–98. [PubMed: 32308832]

17. Metcalfe L, Murrelle EL, Vu L, Joyce AR, Averhart Preston V, Maryon T, et al. Independent Validation in a Large Privately Insured Population of the Risk Index for Serious Prescription Opioid-Induced Respiratory Depression or Overdose. Pain Med. 2020;21(10):2219–28. [PubMed: 32191316]

18. 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. [PubMed: 30901048]

19. Lo-Ciganic WH, Huang JL, Zhang HH, Weiss JC, Kwoh CK, Donohue JM, et al. Using machine learning to predict risk of incident opioid use disorder among fee-for-service Medicare beneficiaries: A prognostic study. PLoS One. 2020;15(7):e0235981. [PubMed: 32678860]

20. Gao W, Leighton C, Chen Y, Jones J, Mistry P. Predicting opioid use disorder and associated risk factors in a Medicaid managed care population. Am J Manag Care. 2021;27(4):148–54. [PubMed: 33877773]

21. 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. [PubMed: 26077738]

22. Vitzthum LK, Riviere P, Sheridan P, Nalawade V, Deka R, Furnish T, et al. Predicting Persistent Opioid Use, Abuse, and Toxicity Among Cancer Survivors. J Natl Cancer Inst. 2020;112(7):720–7. [PubMed: 31754696]

23. Beliveau A, Castilloux AM, Tannenbaum C, Vincent P, de Moura CS, Bernatsky S, et al. Predictors of long-term use of prescription opioids in the community-dwelling population of adults without a cancer diagnosis: a retrospective cohort study. CMAJ Open. 2021;9(1):E96–E106.

24. Ishwaran H, O'Brien R. Commentary: The problem of class imbalance in biomedical data. J Thorac Cardiovasc Surg. 2021;161(6):1940–1. [PubMed: 32711988]

25. Megahed FM, Chen YJ, Megahed A, Ong Y, Altman N, Krzywinski M. The class imbalance problem. Nat Methods. 2021;18(11):1270–2. [PubMed: 34654918]

26. Lyashevska O, Malone F, MacCarthy E, Fiehler J, Buhk JH, Morris L. Class imbalance in gradient boosting classification algorithms: Application to experimental stroke data. Stat Methods Med Res. 2021;30(3):916–25. [PubMed: 33356965]

27. Fulton-Kehoe D, Sullivan MD, Turner JA, Garg RK, Bauer AM, Wickizer TM, et al. Opioid poisonings in Washington State Medicaid: trends, dosing, and guidelines. Med Care. 2015;53(8):679–85. [PubMed: 26172937]

28. He H, Garcia EA. Learning from imbalanced data. IEEE Transactions on knowledge and data engineering. 2009;21(9):1263–84.

29. Cook NR. Use and misuse of the receiver operating characteristic curve in risk prediction. Circulation. 2007;115(7):928–35. [PubMed: 17309939]

30. Szalavitz M The pain was unbeearable. So why did doctors turn her away? WIRED. 2021.

31. Huzienga JE BB, Patel VR, Speights DB. NARxCHECK score as a predictor of unintentional overdose death Appriss Health. 2016.

32. Kilby AE, editor Algorithmic fairness in predicting opioid use disorder using machine learning. Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021.

33. Geissert P, Hallvik S, Van Otterloo J, O'Kane N, Alley L, Carson J, et al. High-risk prescribing and opioid overdose: prospects for prescription drug monitoring program-based proactive alerts. Pain. 2018;159(1):150–6. [PubMed: 28976421]

34. Tseregounis IE, Henry SG. Assessing opioid overdose risk: a review of clinical prediction models utilizing patient-level data. Transl Res. 2021;234:74–87. [PubMed: 33762186]

35. Kuhn M Caret package. Journal of Statistical Software. 2008;28(5).

36. Rufibach K Use of Brier score to assess binary predictions. Journal of clinical epidemiology. 2010;63(8):938–9. [PubMed: 20189763]

37. Cheng A, Kessler D, Mackinnon R, Chang TP, Nadkarni VM, Hunt EA, et al. Reporting guidelines for health care simulation research: Extensions to the CONSORT and STROBE statements. BMJ Simul Technol Enhanc Learn. 2016;2(3):51–60.

38. 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. [PubMed: 25738806]

39. Kuhn M, Johnson K. Applied predictive modeling: Springer; 2013.

40. Lin W-C, Tsai C-F, Hu Y-H, Jhang J-S. Clustering-based undersampling in class-imbalanced data. Information Sciences. 2017;409:17–26.

41. Kang Q, Chen X, Li S, Zhou M. A noise-filtered under-sampling scheme for imbalanced classification. IEEE transactions on cybernetics. 2016;47(12):4263–74. [PubMed: 28113413]

42. Ng WW, Hu J, Yeung DS, Yin S, Roli F. Diversified sensitivity-based undersampling for imbalance classification problems. IEEE transactions on cybernetics. 2014;45(11):2402–12. [PubMed: 25474818]

43. Ramentol E, Caballero Y, Bello R, Herrera F. SMOTE-RSB*: a hybrid preprocessing approach based on oversampling and undersampling for high imbalanced data-sets using SMOTE and rough sets theory. Knowledge and information systems. 2012;33(2):245–65.

44. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. International Journal of Machine Learning and Computing. 2013;3(2):224.

45. Santoso B, Wijayanto H, Notodiputro K, Sartono B, editors. Synthetic over sampling methods for handling class imbalanced problems: A review. IOP conference series: earth and environmental science; 2017: IOP Publishing.

46. Rendon E, Alejo R, Castorena C, Isidro-Ortega FJ, Granda-Gutierrez EE. Data sampling methods to deal with the big data multi-class imbalance problem. Applied Sciences. 2020;10(4):1276.

47. Ling CX, Sheng VS. Cost-sensitive learning and the class imbalance problem. Encyclopedia of machine learning. 2008;2011:231–5.

48. Eubanks V Automating inequality: How high-tech tools profile, police, and punish the poor: St. Martin's Press; 2018.

49. Noble SU. Algorithms of oppression: New York University Press; 2018.

50. Oliva JD. Dosing Discrimination: Regulating PDMP Risk Scores. 2021.

51. Agnoli A, Xing G, Tancredi DJ, Magnan E, Jerant A, Fenton JJ. Association of dose tapering with overdose or mental health crisis among patients prescribed long-term opioids. JAMA. 2021;326(5):411–9. [PubMed: 34342618]

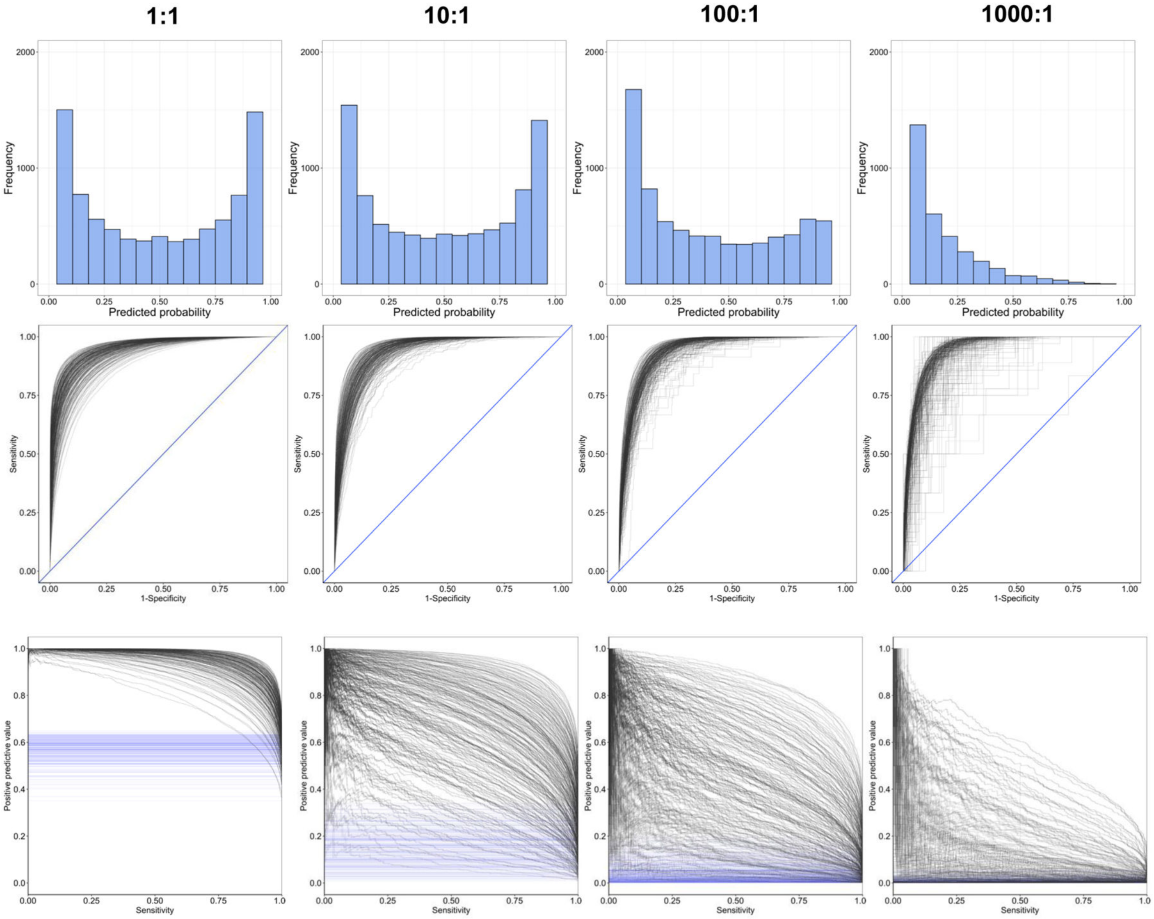

Figure 1.

Histograms of risk scores, receiver operating characteristic (ROC), and precision-recall curves for all models.[a]

[a] Top row: histograms of predicted risk scores. Middle row: receiver operating characteristics (ROC) curves. Bottom row: precision-recall curves. Within each row, the degree of imbalance increases from left to right: 1:1 for the leftmost column, 10:1, 100:1, and 1000:1 for the rightmost column. In the bottom row, the outcome prevalence of each simulation run is shown as a horizontal blue line. Abbreviations: AUC: area under the curve; PRC: area under the precision-recall curve.

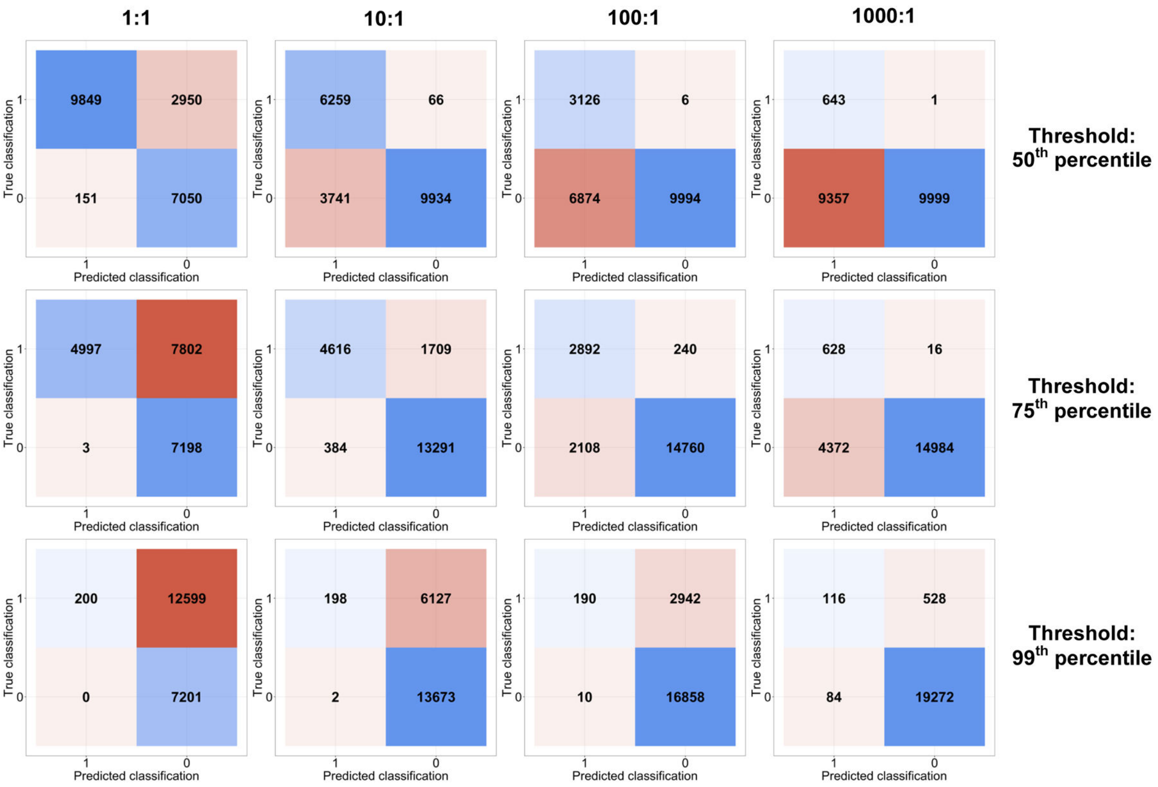

Figure 2.

Confusion matrices showing accuracy of predicted classifications.[a]

[a] Each confusion matrix shows the frequency of observations in each quadrant. Concordant quadrants (where the predictions are correct) are shaded in blue, while discordant quadrants (where predictions are incorrect) are shaded in coral. Quadrants are shaded by frequency such that those with more observations are shaded darker. Clockwise from the top left quadrant: true positives, false negatives, true negatives, false positives.

Table 1.

Outcome, data source, outcome sampling strategy, cumulative incidence of outcome, and approximate imbalance ratio in 20 risk prediction papers for opioid-related outcomes.[a]

First author	Year	Data source	Outcome	Cumulative incidence of outcome	Outcome sampling strategy used	Outcome class ratio	Prediction window	Modeling approach
Zedler *et al.*(13)	2018	Claims	Any opioid overdose or serious opioid-induced respiratory depression	25%	Yes	4:1	4 years	Logistic regression
Dong *et al.*(16)	2019	EHR	Any opioid overdose	10%	Yes	10:1	Variable	Deep learning
Metcalfe *et al.*(17)	2020	Claims	Any opioid overdose or serious opioid-induced respiratory depression	10%	Yes	10:1	2 years	Logistic regression
Zedler *et al.*(21)	2015	Claims	Any opioid overdose or serious opioid-induced respiratory depression	10%	Yes	10:1	2 years	Logistic regression
Béliveau *et al.*(23)	2021	Claims	Long-term opioid use	3.3%	No	100:1	1 year	Logistic regression
Chang *et al.*(7)	2019	PDMP	Nonfatal opioid overdose	3.2%	No	100:1	1 year	Logistic regression
Gao *et al.*(20)	2021	Claims	Incident opioid use disorder	2.0%	No	100:1	1 year	Logistic regression
Lo-Ciganic *et al.*(19)	2020	Claims	Incident opioid use disorder	1.54%	No	100:1	Variable	Machine learning (various)
Vitzthum *et al.*(22)	2020	Claims	Incident opioid use disorder	2.9%	No	100:1	15 years	LASSO regression
Cochran *et al.*(15)	2014	Claims	Incident opioid use disorder	0.10%	No	1000:1	2 years	Logistic regression
Dong *et al.*(12)	2021	EHR	Any opioid overdose	0.86%	No	1000:1	Variable	Sequential deep learning
Ellis *et al.*(11)	2019	EHR	Opioid dependence and overdose	0.90%	No	1000:1	Variable	Random forest
Glanz *et al.*(10)	2018	Claims	Fatal prescription and heroin-related overdose	0.50%	No	1000:1	2 years	Cox proportional hazards
Ferris *et al.*(6)	2021	PDMP	Fatal opioid overdose	0.14%	No	1000:1	2 years	Logistic regression
Lo-Ciganic *et al.*(18)	2019	Claims	Fatal and nonfatal opioid overdose	0.57%	No	1000:1	Variable	Machine learning (various)
Saloner *et al.*(9)	2020	PDMP	Nonfatal opioid overdose	0.37%	No	1000:1	1 year	Logistic regression
Chang *et al.*(8)	2020	PDMP	Fatal opioid overdose	0.07%	No	10,000:1	3, 6, 9, 12 months	Logistic regression
Ferris *et al.*(5)	2019	PDMP	Any opioid overdose	0.09%	No	10,000:1	8 months	Logistic regression
Geissert *et al.*(33)	2018	PDMP	Fatal prescription overdose	0.02%	No	10,000:1	1 year	Logistic regression
Sun *et al.*(14)	2020	EHR	Any opioid overdose	0.05%	No	10,000:1	30 days	Machine learning (various)

[a] Abbreviations. PDMP: Prescription drug monitoring program; EHR: electronic health record; VA: Veterans' Administration. Where papers report multiple outcomes, we have chosen just one outcome to present in this table. The last column in the table ("outcome sampling strategy used") assesses whether each paper employed some kind of sampling strategy that effectively reduced the degree of imbalance between non-cases and cases.

Table 2.

Calibration and predictive performance of logistic regression and random forest models on balanced and imbalanced data.

Risk score cutoff (percentile)[a]	Model	Imbalance ratio	Brier score	Sensitivity	Specificity	Positive predictive value	Negative predictive value	Overall accuracy
50th	SuperLearner	1:1	0.09	0.81	0.91	0.91	0.79	0.85
	SuperLearner	10:1	0.06	0.99	0.59	0.31	0.99	0.65
	SuperLearner	100:1	0.03	0.99	0.53	0.09	0.99	0.54
	SuperLearner	1000:1	0.005	0.99	0.50	0.01	0.99	0.51
75th	SuperLearner	1:1	0.09	0.45	0.99	0.97	0.58	0.68
	SuperLearner	10:1	0.06	0.86	0.86	0.51	0.97	0.85
	SuperLearner	100:1	0.03	0.94	0.79	0.18	0.97	0.79
	SuperLearner	1000:1	0.005	0.95	0.75	0.02	0.99	0.76
99th	SuperLearner	1:1	0.09	0.02	0.99	0.99	0.44	0.45
	SuperLearner	10:1	0.06	0.07	0.99	0.85	0.85	0.85
	SuperLearner	100:1	0.03	0.15	0.99	0.53	0.96	0.95
	SuperLearner	1000:1	0.005	0.22	0.99	0.14	0.99	0.99

[a] This refers to the threshold, or percentile of the risk score distribution, at which a risk score is classified as a case or a non-case. For example, if the 75th percentile is used, the 75th percentile of the risk score distribution generated by the logistic regression model is identified. Any observations with a risk score greater than the 75th percentile is classified as a case (1), while any with a risk score equal to or lower than the 75th percentile is classified as a non-case. This process is repeated for the random forest model and for other classification threshold