BMJ Health & Care Informatics

Resampling to address inequities in predictive modeling of suicide deaths

Majerle Reeves ,[1] Harish S Bhat,[1] Sidra Goldman-Mellor[2]

To cite: Reeves M, Bhat HS, Goldman-Mellor S. Resampling to address inequities in predictive modeling of suicide deaths. *BMJ Health Care Inform* 2022;**29**:e100456. doi:10.1136/bmjhci-2021-100456

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjhci-2021-100456).

Received 31 July 2021
Accepted 03 March 2022

[1]Department of Applied Mathematics, University of California Merced, Merced, California, USA
[2]Department of Public Health, University of California Merced, Merced, California, USA

Correspondence to
Majerle Reeves;
mreeves3@ucmerced.edu

ABSTRACT

Objective Improve methodology for equitable suicide death prediction when using sensitive predictors, such as race/ethnicity, for machine learning and statistical methods.

Methods Train predictive models, logistic regression, naive Bayes, gradient boosting (XGBoost) and random forests, using three resampling techniques (Blind, Separate, Equity) on emergency department (ED) administrative patient records. The Blind method resamples without considering racial/ethnic group. Comparatively, the Separate method trains disjoint models for each group and the Equity method builds a training set that is balanced both by racial/ethnic group and by class.

Results Using the Blind method, performance range of the models' sensitivity for predicting suicide death between racial/ethnic groups (a measure of prediction inequity) was 0.47 for logistic regression, 0.37 for naive Bayes, 0.56 for XGBoost and 0.58 for random forest. By building separate models for different racial/ethnic groups or using the equity method on the training set, we decreased the range in performance to 0.16, 0.13, 0.19, 0.20 with Separate method, and 0.14, 0.12, 0.24, 0.13 for Equity method, respectively. XGBoost had the highest overall area under the curve (AUC), ranging from 0.69 to 0.79.

Discussion We increased performance equity between different racial/ethnic groups and show that imbalanced training sets lead to models with poor predictive equity. These methods have comparable AUC scores to other work in the field, using only single ED administrative record data.

Conclusion We propose two methods to improve equity of suicide death prediction among different racial/ethnic groups. These methods may be applied to other sensitive characteristics to improve equity in machine learning with healthcare applications.

Summary

What is already known?

- There has been significant research in building machine learning/statistical models for predicting suicide.
- Most of these models use race as a predictor, but do not include analysis of how this predictor is used.
- Most of these models follow patients over a period of time and do not analyse a single visit.

What does this paper add?

- Shows models can perform competitively only using one patient visit and administrative patient records.
- Compares model performance on different racial/ethnic groups.
- Introduces two resampling techniques to increase racial/ethnic equitability in model performance.

INTRODUCTION

Suicide is the 10th leading cause of death in the USA and has increased 35% from 1999 to 2018.[1] Despite decades of clinical and epidemiological research, our ability to predict which individuals will die by suicide has not improved significantly in the last 50 years.[2] Many factors (eg, prior non-fatal suicide attempt, psychiatric disorder, stressful life events and key demographic characteristics) are associated with elevated suicide risk at the population level, but individualised suicide risk prediction remains challenging.

Recent research attempting to improve the performance of previous suicide prediction models has used statistical and machine learning tools to explore suicide risk factors and to classify patients according to their risk for suicidal behaviour.[3–9] Much of this work has focused on patients in healthcare settings, motivated by the growing availability of large-scale longitudinal health data through electronic medical record (EMR) systems, the high proportion of suicide decedents who have contact with healthcare providers in the year before their deaths,[10] and healthcare patients' substantially elevated risks of suicide.[11] Many of these studies focus on high-risk groups[5 6 9] and/or predicting non-fatal suicidal behaviours[7 8] instead of suicide death, due to the low base rate of suicide and/or the difficulty of linking EMRs with death records.

The increasing prominence of machine learning models in healthcare applications has been accompanied by increasing concerns that these models perpetuate and potentially exacerbate long-standing inequities in the provision and quality of healthcare services.[12 13] Algorithmic unfairness can

stem from two places: the collected data and the machine learning algorithms.[14] To address this issue, several groups[15 16] have advocated for machine learning models to be proactively designed in ways that advance equity in health outcomes and prioritise fairness. This goal is critical in the mental healthcare and suicide prevention fields, where research has long documented both racial discrimination in care as well as racial/ethnic disparities in rates of suicidal behaviour and mental health stigma.[17–19] Recent work has shown that predictive models for suicide death are less accurate for Native American/Alaskan Aleut, non-Hispanic Black and patients with unknown racial/ethnic information compared with Hispanic, non-Hispanic White or Asian patients.[9] Although the ultimate goal is ensuring that racial/ethnic minoritised groups derive equal benefit with respect to patient outcomes from the deployment of machine learning models in healthcare systems, an important goal in the earlier stages of model development is testing whether a prediction model is equally accurate for patients in minoritised and non-minority groups.[15 20]

We build models that quantify an individual's risk of future death by suicide, using information gleaned from a single visit to an emergency department to seek care for any condition, including non-psychiatric conditions. Our retrospective cohort study uses a database of administrative patient records (APRs) linked with death records that has not been used in prior predictive modelling studies. To address the low base rate of suicide death and/or racial/ethnic imbalances, we resample database records to build three different training sets. Using metrics established in the literature, we measure the test set performance of four classifiers trained on each of the three resampled training sets, focusing on methods that equalise opportunity and odds across all subgroups.

METHODS

Data sources

This study uses APRs provided by the California Office of Statewide Health Planning and Development together with linked death records provided by the California Department of Public Health Vital Records. All data obtained and used were deidentified.

We analyse all visits to all California-licensed EDs from 2009 to 2012, by individuals aged at least five with a California residential zip code and less than 500 visits. The data contains N=35 393 415 records from 12 818 456 patients,[21] and includes the date and underlying cause of death for all decedents who died in California in 2009–2013.

For each record, we assign a label of Y=1 if the record corresponds to a patient who died by suicide (corresponding to International Classification of Diseases-version 10 (ICD-10) codes X60-X84, Y87.0 or U03) during the period 2009–2013; otherwise, we assign a label of Y=0. This allows a minimum of 1 year between each patient visit and when deaths are assessed. The goal of our models is to use information from a single visit by a single patient to predict Y, death by suicide between 2009 and 2013. In our records, 9364 patients (with 37 661 records) died by suicide; as <0.11% of the data is in the Y=1 (death by suicide) class, the classification problem is imbalanced.

The APRs contain both patient- and facility-level information which includes basic patient demographic characteristics, insurance/payer status, discharge information, type of care, admission type and one primary and up to five secondary Clinical Classifications Software (CCS) diagnostic codes. These CCS codes aggregate more than 14 000 ICD-9-CM diagnoses into 285 mutually exclusive and interpretable category codes. The APRs also contain supplemental E-Codes, which provide information about the intent (accidental, intentional, assault, or undetermined) of external injuries and poisonings. Note that APRs omit information such as vital signs, height/weight and other biological indicators found in a full medical record. See online supplemental material for additional information regarding APRs.

Table 1 breaks down the data set by racial/ethnic identity. Seven categories describe racial/ethnic identity: Black, Native American/Eskimo/Aleut, Asian/Pacific Islander, White, unknown/invalid/blank, other and Hispanic. While we recognise that these are crude measures for racial/ethnic identity, this is the granularity of information collected by hospitals and used in machine learning models. Native American/Eskimo/Aleut and White patients have significantly higher rates of suicide death than Hispanic, Black or Asian/Pacific Islander patients, which is consistent with the measured trends.[1] In this work, we do not train classification models for the Native American/Eskimo/Aleut group, as the number of suicide deaths is too small to generalise to a wider population. Note that racial/ethnic

Table 1 Data broken down by race/ethnic feature, excluding the 'other' and 'unknown' race categories

	White	Hispanic	Black	Asian	Native American
Patient records	17 337 370	9 863 670	4 437 649	2 014 810	125 769
Suicide death records	27 974	4739	2099	1246	264
Suicide death records per 100 000 patient records	161	48	47	62	210

Note that suicide rates differ considerably by category.

information is supposed to be self-reported by patients but may be inferred incorrectly by clinical personnel or be incorrectly recorded[22]; we assume the error rate is low enough to not affect our results substantially.

Statistical methods

Given the large imbalance in the class distribution, training directly on the raw data would yield classifiers that achieve accuracies exceeding 99.9% by predicting that no one dies by suicide. To derive meaningful results, we must proactively address the class imbalance; we focus on resampling, an established approach for classification with imbalanced data.[23]

For each of three resampling methods (denoted below as Blind, Separate and Equity), we apply four statistical/machine learning techniques: logistic regression, naive Bayes, random forests and gradient boosted trees (model descriptions in online supplemental material). This yields 12 models, which we compare below. In each case, we split the raw data into training, validation, and test sets and resample only the training sets. We select model hyperparameters (eg, for tree-based models, the maximum depth of the tree) by assessing the performance of trained models on validation sets. Once we have selected hyperparameters and finished training a model, we report its test set performance.[24] The test set is not used for any other purpose, simulating a scenario in which a model is applied to newly collected data.

For imbalanced binary classification problems, among the most widely used resampling methods are those that sample uniformly from either or both classes to create a class-balanced training set.[23] We choose this method as a baseline and denote this method as Blind; it resamples without considering racial/ethnic group membership. The Separate and Equity resampling procedures are different ways to account for racial/ethnic group membership when forming balanced training sets. These sampling techniques address two sources of bias in the data: representation bias and aggregation bias. From table 1, the White population comprises the majority of patient records as well as suicide deaths, leaving all minority groups underrepresented. The aggregation of over-represented data with underrepresented data can lead to bias. However, there can still be aggregation bias when groups are equally represented.[14] For this reason, we train separate models for each racial/ethnic group in addition to a joint model with Equity resampling.

For all three approaches, we begin by shuffling the data by unique patient identifier. We then divide the data into training, validation, and test sets with a roughly 60/20/20 ratio, ensuring that each set is disjoint in terms of patients. This ensures that patients used for training are not in the test set, which may artificially inflate model performance.

In the Blind method, we separate the training set by class, resulting in two sets. We then randomly sample a subset of the majority class—patients who do not die of suicide—till we achieve a balanced training set.

In the Separate method, the training data is separated by racial/ethnic group and like the Blind method we undersample the majority class to balance the data. We thus train disjoint models for each racial/ethnic group.

In contrast, in the Equity method, we divide the training set by both racial/ethnic group and class label. This results in eight training subsets. We then sample 7500 files with replacement from each of the 8 training subsets. The union of these samples is the equity-directed resampled training set; note its balance across racial/ethnic groups and across 0/1 labels. This is a form of stratified resampling in which the strata are racial/ethnic group and 0/1 label.[25] In this case, the trained model can be applied to test data from any of the four groups.

In these models, we treat each visit by each patient independently. Consequently, each predictive model bases its prediction only on the APR from the current (index) visit. As resampling uses randomness, we show the robustness of our results by repeating the sampling procedure and building/training the models with 10 different random seeds. Additional information about the random trials can be found in online supplemental material, figures S1-S3. When reporting the results, we provide average performance (with SD) of each model for each racial/ethnic group and resampling method.

RESULTS

In tables 2–4, we report test set sensitivity, specificity and area under the curve (AUC)[24] for each resampling method and model type, broken down by racial/ethnic group. We do not report accuracy due to the class imbalance. Here sensitivity and specificity are, respectively, the percentages of correctly classified records in the Y=1 (patients who died by suicide) and Y=0 (all other patients) classes. When analysing the performance of different models, we imagine a setting in which patients classified as positive (ie, at high risk of suicide) have the opportunity to receive an intervention such as a postdischarge phone call. We, thus, prioritise sensitivity over specificity, as false negatives consist of patients who die of suicide with no intervention, while false positives consist of patients who receive a potentially unneeded intervention.

Our models output a probability of Y=1 (suicide death) conditional on a patient's APR. In each case, there is a threshold τ such that when the model output exceeds (respectively, does not exceed) τ, we assign a predicted label of 1 (respectively, 0).[26] We assign τ values to each model to approximately balance specificity, enabling comparison of models based on test set sensitivity. We also report the size of range which is the difference between the highest performance and lowest performance by racial/ethnic group. A smaller range implies more equitable performance across the racial/ethnic groups; a model whose range is zero (for both sensitivity and

Table 2 Average test set sensitivity with SD (at training set specificity of approximately 0.76) of different combinations of resampling procedure plus statistical/machine learning method, by racial/ethnic group

	Asian	Black	Hispanic	White	Size of range
Blind—Logistic Regression	0.43 (0.05)	0.30 (0.08)	0.32 (0.04)	0.76 (0.02)	0.47 (0.05)
Blind—Naive Bayes	**0.44** (0.05)	**0.35** (0.09)	**0.38** (0.04)	0.70 (0.02)	**0.37** (0.05)
Blind—XGBoost	0.37 (0.03)	0.27 (0.08)	0.30 (0.03)	**0.81** (0.03)	0.56 (0.05)
Blind—Random Forest	0.31 (0.03)	0.24 (0.07)	0.25 (0.03)	0.79 (0.04)	0.58 (0.05)
Separate—Logistic Regression	0.69 (0.03)	0.56 (0.09)	0.63 (0.04)	**0.58** (0.02)	0.16 (0.07)
Separate—Naive Bayes	0.60 (0.04)	**0.61** (0.11)	0.57 (0.04)	0.56 (0.03)	**0.13** (0.06)
Separate—XGBoost	0.67 (0.04)	0.53 (0.12)	0.57 (0.07)	**0.58** (0.02)	0.19 (0.09)
Separate—Random Forest	**0.71** (0.04)	**0.61** (0.12)	**0.67** (0.05)	0.57 (0.03)	0.20 (0.08)
Equity—Logistic Regression	0.58 (0.03)	0.52 (0.09)	0.63 (0.04)	0.61 (0.02)	0.14 (0.05)
Equity—Naive Bayes	0.56 (0.05)	0.57 (0.09)	0.63 (0.05)	0.59 (0.03)	0.12 (0.04)
Equity—XGBoost	0.55 (0.04)	0.50 (0.10)	0.69 (0.03)	**0.70** (0.02)	0.24 (0.06)
Equity—Random Forest	**0.65** (0.08)	**0.65** (0.11)	**0.72** (0.07)	**0.70** (0.06)	**0.13** (0.07)

For each column and each resampling method, boldface indicates the top performing method(s).

specificity) satisfies the equal odds criterion established in the algorithmic fairness literature.

We see several trends in the results. First, Blind resampling is the least equitable in terms of either test set sensitivity or test set specificity. All models yield worse test sensitivity on minoritised racial/ethnic groups than on the White group. Models trained with Blind resampling learn to overclassify White patient files as dying of suicide. The AUC metric obscures these differences and makes them hard to discern. These results hold for all four statistical/machine learning methods considered.

Both the Separate and Equity resampling strategies lead to more equalised sensitivity and specificity across the four racial/ethnic groups. These strategies lead all four statistical/machine learning methods to improve in terms of the equal odds criteria for fairness in classification[20] and treatment equality; these strategies reduce the range in performance of false negative and false positive rates across the different racial/ethnic groups.[14 27] For instance, the range of sensitivities for logistic regression decreases from 0.47 (Blind) to either 0.16 (Separate) or 0.14 (Equity). Notably, this reduction in performance range is coupled with a boost in test set sensitivities on the minority racial/ethnic groups, and a boost in test set specificity for the White group. For further discussion on fairness, see online supplemental material.

Table 4 shows that the test set AUC of XGBoost (with Equity resampling) is between 0.73 and 0.78, signifying good diagnostic accuracy.[26] This is clearly better than random guessing (AUC of 0.5) and exceeds all AUC scores reported in a meta-analysis of 50 years of suicide modelling.[2] Our AUC scores are comparable to a recent study's male-specific models (0.77 for CART[28] and 0.80 for random forests), and slightly less than that study's

Table 3 Average test set specificity with SD (at training set specificity of approximately 0.76) of different combinations of resampling procedure plus statistical/machine learning method, by racial/ethnic group

	Asian	Black	Hispanic	White	Size of range
Blind—Logistic Regression	0.91 (0.01)	0.93 (0.01)	0.94 (0.01)	0.57 (0.01)	0.38 (0.01)
Blind—Naive Bayes	0.91 (0.01)	0.90 (0.01)	0.91 (0.00)	**0.61** (0.01)	**0.30** (0.01)
Blind—XGBoost	0.96 (0.00)	0.95 (0.01)	0.95 (0.00)	0.54 (0.02)	0.42 (0.02)
Blind—Random Forest	**0.97** (0.00)	**0.96** (0.00)	**0.96** (0.00)	0.52 (0.04)	0.45 (0.04)
Separate—Logistic Regression	0.66 (0.01)	0.66 (0.02)	0.74 (0.01)	0.75 (0.00)	**0.10** (0.01)
Separate—Naive Bayes	**0.72** (0.03)	0.63 (0.08)	**0.78** (0.01)	0.74 (0.01)	0.15 (0.07)
Separate—XGBoost	0.71 (0.02)	**0.73** (0.06)	0.76 (0.05)	**0.77** (0.01)	**0.10** (0.04)
Separate—Random Forest	0.61 (0.03)	0.64 (0.03)	0.70 (0.03)	0.75 (0.02)	0.15 (0.04)
Equity—Logistic Regression	0.79 (0.01)	0.78 (0.01)	**0.76** (0.01)	**0.71** (0.01)	**0.08** (0.01)
Equity—Naive Bayes	**0.81** (0.01)	0.74 (0.01)	**0.76** (0.02)	**0.71** (0.02)	0.10 (0.01)
Equity—XGBoost	**0.81** (0.01)	**0.80** (0.02)	0.72 (0.01)	0.65 (0.01)	0.17 (0.01)
Equity—Random Forest	0.70 (0.06)	0.65 (0.03)	0.66 (0.04)	0.61 (0.06)	0.10 (0.02)

For each column and each resampling method, boldface indicates the top performing method(s).

Table 4 Average test set AUC with SD (at training set specificity of approximately 0.76) of different combinations of resampling procedure plus statistical/machine learning method, by racial/ethnic group

	Asian	Black	Hispanic	White	Size of range
Blind—Logistic Regression	0.77 (0.02)	**0.73** (0.05)	0.77 (0.02)	0.73 (0.01)	**0.06** (0.04)
Blind—Naive Bayes	0.74 (0.03)	**0.73** (0.04)	0.75 (0.02)	0.71 (0.01)	**0.06** (0.03)
Blind—XGBoost	**0.79** (0.02)	**0.73** (0.05)	**0.79** (0.01)	**0.76** (0.01)	0.07 (0.05)
Blind—Random Forest	0.77 (0.02)	0.72 (0.05)	0.76 (0.01)	0.73 (0.01)	0.07 (0.04)
Separate—Logistic Regression	0.74 (0.02)	0.66 (0.05)	**0.75** (0.02)	0.73 (0.01)	0.10 (0.05)
Separate—Naive Bayes	0.73 (0.03)	0.67 (0.05)	0.74 (0.02)	0.71 (0.01)	**0.09** (0.04)
Separate—XGBoost	**0.76** (0.02)	**0.69** (0.06)	**0.75** (0.02)	**0.75** (0.01)	**0.09** (0.05)
Separate—Random Forest	0.74 (0.02)	0.67 (0.07)	**0.75** (0.01)	0.73 (0.01)	0.10 (0.06)
Equity—Logistic Regression	0.76 (0.02)	0.71 (0.05)	0.77 (0.02)	0.72 (0.01)	0.08 (0.04)
Equity—Naive Bayes	0.74 (0.03)	0.72 (0.05)	0.76 (0.02)	0.71 (0.01)	**0.07** (0.03)
Equity—XGBoost	**0.77** (0.03)	**0.73** (0.06)	**0.78** (0.01)	**0.74** (0.01)	0.08 (0.05)
Equity—Random Forest	0.76 (0.03)	0.70 (0.06)	0.76 (0.01)	0.72 (0.01)	0.09 (0.05)

For each column and each resampling method, boldface indicates the top performing method(s).
AUC, area under the curve.

female-specific models (0.87 for CART and 0.88 for random forests).[4]

DISCUSSION

We trained machine learning models on statewide emergency department using three resampling methods on APRs for suicide death classification. We have shown that resampling methods can reduce the range in model performance on different racial/ethnic groups by at least 50%. Specifically, equity-focused resampling increases the predictive performance of all four machine learning models on minoritised racial/ethnic patient groups to approximately match that of the majority (non-Hispanic White) patient group.

This study has several strengths. Our models achieve high predictive accuracy using only single-visit APRs, whereas other studies often have a much richer feature space from which to learn[4] and/or restrict attention to only those patients with at least three visits.[8] Additionally, the resampling and machine learning methods we employ are highly scalable. Given additional records from other healthcare systems (eg, from neighbouring states), we could add them to our current data set and resample/retrain without difficulty. Our models also issue predictions for the general population of emergency department patients instead of subpopulations with higher suicide risk,[5 6] increasing their scope and generalisability. We also use linked mortality records to predict suicide death rather than nonfatal suicidal behaviours or self-harm.[3 7 8] When previous large-scale machine learning models have been trained on such linked data sets, they have often used data from nationalised/centralised systems unavailable in the USA.[4] Additionally, the Equity method allows for learning across racial/ethnic groups, but because each racial/ethnic group is equally represented, still allows for racial/ethnic specific predictors to be identified. For example, a mental health diagnosis is a recognised predictor for suicide death,[2] but non-Hispanic Black, Hispanic and Asian individuals are less likely to be diagnosed with a mental health condition.[18 19]

While we may be able to improve on our methods with additional features such as lab results and medication history, there are benefits to training with APRs. The features in APRs are accessible in most (if not all) existing EMR databases. Because these models are trained solely on information gathered at a single emergency department visit, there is no need to process a patient's medical history. While the logistic regression and naive Bayes models are inherently interpretable, the boosted tree and random forest models could also be analysed and interpreted in detail prior to implementation. Therefore, deployment of these methods as an extension of existing database software is feasible. We envision this as a tool that could potentially assist healthcare providers in identifying patients at risk for suicide death.[29 30]

Other machine learning for healthcare applications can benefit from this equity analysis. We showed that regardless of model type, the Blind resampling method resulted in inequitable suicide classification for different racial/ethnic groups. Our findings suggest that sensitive group representation should be considered as a type of class imbalance that must be rectified before model training takes place. While we have focused here on racial/ethnic group membership, the Separate or Equity resampling methods can be directly applied to other sensitive categories. We hypothesise that in other problem domains and applications, one can improve prediction equity either by building separate models, or by using equity-directed resampling. When separation of data by sensitive group results in sample sizes too small to train machine learning models, equity-directed resampling may still be viable.

This study also has limitations. First, as with all machine learning models, the finalised predictions are intended to complement (rather than substitute for) human judgement. As with other technology (eg, medical imaging), practitioners may require additional explanation/interpretation of what the models do internally, to trust and apply their predictions in a beneficial way. Though we address algorithmic fairness, we should not expect purely technological solutions to address systemic inequities in the healthcare system.[13] These inequities may cause unequal mislabeling of suicides by race/ethnicity, affecting the quality of the linked data we analyse, and thereby reducing the true generalisability of our models to real-world settings.[18] Within the algorithmic fairness context, while our equity-resampled models achieve predictive equality across racial/ethnic groups, recorded membership in these groups is not always accurate. Additionally, patients potentially belong to many vulnerable groups (via their socioeconomic status, disability status, Veteran status, etc); further resampling/stratification may be needed to achieve algorithmic fairness with respect to all such groups. In some cases, for instance, if sample sizes are too small, achieving the equal opportunity standard may not be possible. Finally, because we have trained our models only on data from California residents in specific years, we cannot be sure that the trained models themselves will generalise to other locations and time periods. However, the techniques we describe could be applied to construct analogous models given sufficient data from other locations.

CONCLUSION

When building suicide prediction models using highly imbalanced data sets, resampling is necessary. However, blind resampling can negatively impact model performance for minority groups. Applying either of two resampling methods, we develop predictive models that have reduced prediction inequity across racial/ethnic groups.

Contributors MR is the primary author of the manuscript, with contributions and editing from HSB and SG-M. MR conducted all statistical and machine learning analyses; HSB provided guidance on construction and application of these methods. MR and HSB conceived of the study. SG-M obtained the data and gave feedback on results and their interpretation. HSB is the guarantor.

Funding This project was funded through the University of California Firearm Violence Research Center and through National Institute of Mental Health grant R15 MH113108-01 to SG-M. MR acknowledges NRT Fellowship support from National Science Foundation grant DGE-1633722. We acknowledge computational support from the MERCED cluster, supported by National Science Foundation grant ACI-1429783.

Disclaimer The sponsors had no role in the study design; collection, analysis, or interpretation of data; writing of the report, or decision to submit the article for publication.

Competing interests None declared.

Patient consent for publication Not applicable.

Ethics approval This study was approved by Institutional Review Boards of the California Health and Human Services Agency (17-02-2890) and the University of California, Merced (UCM2017-154) and follows relevant TRIPOD guidelines.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available. Not applicable.

ORCID iD

Majerle Reeves http://orcid.org/0000-0002-2112-1712

REFERENCES

1 National Institute of Mental Health. Suicide, 2021. Available: https://www.nimh.nih.gov/health/statistics/suicide.shtml [Accessed 13 Jul 2021].
2 Franklin JC, Ribeiro JD, Fox KR, *et al*. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull* 2017;143:187–232.
3 Bhat H, Goldman-Mellor S. Predicting adolescent suicide attempts with neural networks. NIPS 2017 Workshop on Machine Learning for Health (ML4H), 2017. Available: 10057.http://arxiv.org/abs/1711.10057
4 Gradus JL, Rosellini AJ, Horváth-Puhó E, *et al*. Prediction of sex-specific suicide risk using machine learning and single-payer health care registry data from Denmark. *JAMA Psychiatry* 2020;77:25–34.
5 Katz C, Randall JR, Sareen J, *et al*. Predicting suicide with the SAD PERSONS scale. *Depress Anxiety* 2017;34:809–16.
6 Larkin C, Di Blasi Z, Arensman E. Risk factors for repetition of self-harm: a systematic review of prospective hospital-based studies. *PLoS One* 2014;9:e84282.
7 Jung JS, Park SJ, Kim EY, *et al*. Prediction models for high risk of suicide in Korean adolescents using machine learning techniques. *PLoS One* 2019;14:e0217639.
8 Barak-Corren Y, Castro VM, Javitt S, *et al*. Predicting suicidal behavior from longitudinal electronic health records. *Am J Psychiatry* 2017;174:154–62.
9 Coley RY, Johnson E, Simon GE, *et al*. Racial/Ethnic disparities in the performance of prediction models for death by suicide after mental health visits. *JAMA Psychiatry* 2021;78:726.
10 Chock MM, Bommersbach TJ, Geske JL, *et al*. Patterns of health care usage in the year before suicide: a population-based case-control study. *Mayo Clin Proc* 2015;90:1475–81.
11 Goldman-Mellor S, Olfson M, Lidon-Moyano C, *et al*. Association of suicide and other mortality with emergency department presentation. *JAMA Netw Open* 2019;2:e1917571.
12 Gianfrancesco MA, Tamang S, Yazdany J, *et al*. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544–7.
13 McCradden MD, Joshi S, Mazwi M, *et al*. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit Health* 2020;2:e221–3.
14 Mehrabi N, Morstatter F, Saxena N. A survey on bias and fairness in machine learning. *arXiv* 2019;190809635.
15 Rajkomar A, Hardt M, Howell MD, *et al*. Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018;169:866–72.
16 DeCamp M, Lindvall C. Latent bias and the implementation of artificial intelligence in medicine. *J Am Med Inform Assoc* 2020;27:2020–3.
17 Wong EC, Collins RL, McBain RK, *et al*. Racial-Ethnic differences in mental health stigma and changes over the course of a statewide campaign. *Psychiatr Serv* 2021;72:514–20.
18 Rockett IRH, Wang S, Stack S, *et al*. Race/ethnicity and potential suicide misclassification: window on a minority suicide paradox? *BMC Psychiatry* 2010;10:35.
19 Primm AB, Vasquez MJT, Mays RA, *et al*. The role of public health in addressing racial and ethnic disparities in mental health and mental illness. *Prev Chronic Dis* 2010;7:A20.

20 Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems*, 2016: 3315–23.
21 Goldman-Mellor S, Hall C, Cerdá M, *et al*. Firearm suicide mortality among emergency department patients with physical health problems. *Ann Epidemiol* 2021;54:38–44.
22 Saunders CL, Abel GA, El Turabi A, *et al*. Accuracy of routinely recorded ethnic group information compared with self-reported ethnicity: evidence from the English Cancer Patient Experience survey. *BMJ Open* 2013;3. doi:10.1136/bmjopen-2013-002882. [Epub ahead of print: 28 Jun 2013].
23 Kuhn M, Johnson K. *Applied Predictive Modeling*. New York, NY: Springer-Verlag, 2013.
24 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. 2 edn. New York, NY, USA: Springer, 2009.
25 Davison AC, Hinkley DV. *Bootstrap methods and their application*. Cambridge University Press, 1997.
26 Šimundić A-M. Measures of diagnostic accuracy: basic definitions. *EJIFCC* 2009;19:203.
27 Ibrahim SA, Charlson ME, Neill DB. Big data analytics and the struggle for equity in health care: the promise and perils. *Health Equity* 2020;4:99–101.
28 Breiman Leo. *Classification and Regression Trees*. Wadsworth International Group, 1984. Available: http://cds.cern.ch/record/2253780 [Accessed 8 Sep 2019].
29 Belsher BE, Smolenski DJ, Pruitt LD, *et al*. Prediction models for suicide attempts and deaths: a systematic review and simulation. *JAMA Psychiatry* 2019;76:642–51.
30 Simon GE, Matarazzo BB, Walsh CG, *et al*. Reconciling statistical and clinicians' predictions of suicide risk. *Psychiatr Serv* 2021;72:555–62.

Supplementary Online Content

Administrative Patient Records (APRs)

The APRs contain information for each patient at each visit. In this data set, each APR is complete, meaning this is a complete-case analysis. Each record includes 612 predictors. The general categories from which all predictors stem are provided in Table S1.

Table S1. All predictors used for building predictive models. Categorical variables were one hot encoded, generating a space of 612 predictors. ED denotes emergency department while PD denotes inpatient hospitalization.

Description	**Type**	**Description**	**Type**
Age	Numeric	Insurance Category	Categorical (5 levels)
ED Visit	Binary	Disposition (ED)	Categorical (34 levels)
PD Visit	Binary	Disposition (PD)	Categorical (14 levels)
Facility ID Number	Numeric	Payer (ED)	Categorical (22 levels)
Facility Zip	Numeric	Facility County	Categorical (58 levels)
Corrected Zip	Numeric	Type of Care	Categorical (6 levels)
Hospital Zip	Numeric	Source Site	Categorical (10 levels)
Rural / Urban Score	Numeric	Admission Type	Categorial (5 Levels)
Length of Stay	Numeric	Payer Category	Categorial (10 Levels)
Pay Plan	Numeric	Payer Type	Categorial (4 Levels)
Domain Expert Features	One numeric and 19 binary	Patient County	Categorial (58 Levels)
Present on Arrival	Binary	Hospital County	Categorial (58 Levels)
Sex	Categorial (4 Levels)	CCS Diagnostic Code	Categorial (262 Levels)
Race	Categorical (7 levels)	E-Codes	Categorial (24 Levels)

Table S2 contains selected feature prevalence for each of the resampling methods. We have selected a subset of features to display for purposes of illustration. We show in Table S2 the average patient age in each data set, the percentage of files associated with male patients, and the self-harm, suicidal ideation, and CCS codes for mental health and substance abuse per 100,000 files. Note that each data set varies in size due to the resampling rate. As we only resample the training set, All Data represents the distribution for approximately 60% of the raw data, with the remaining data split between the validation and test sets. As such, the validation and test sets will have distributions similar to All Data or data collected in the real world. We utilize the training and validation sets for model development and then use the test set for model validation. Recall that the set of patients in the test set is totally disjoint from the sets of patients in either the validation or training sets. Hence this model is best described as Type 2a under the TRIPOD statement list of prediction types[1].

Reeves M, *et al. BMJ Health Care Inform* 2022; 29:e100456. doi: 10.1136/bmjhci-2021-100456

Table S2. Feature prevalence for each resampling method. All data focuses on the entirety of the training set. We look at average age, sex, self-harm, suicidal ideation, and the CCS codes for mental health and substance abuse (MHSA). This table only describes distributions for one random shuffling, splitting, and sampling of the data.

	All Data		***Blind Resampling***		***Equity Resampling***		***Separate Resampling***							
							Asian		**Black**		**Hispanic**		**White**	
	Suicide	**Non-Suicide**	**Suicide**	**Non-Suicide**	**Suicide**	**Non-Suicide**	**Suicide**	**Non-Suicide**	**Suicide**	**Non-Suicide**	**Suicide**	**Non-Suicide**	**Suicide**	**Non-Suicide**
Average Age	47	46	47	46	44	46	51	53	39	42	39	40	49	51
Men (%)	66	44	66	43	68	43	69	44	73	41	66	45	64	45
Self-Harm per 100,000	6800	380	6800	349	8377	370	10349	134	7291	307	7593	259	6323	401
Suicidal Ideation per 100,000	3505	513	3505	484	3663	440	2957	269	5909	460	3333	481	3419	599
650 MHSA: Adjustment disorders per 100,000	488	161	488	197	867	170	1210	0	1381	384	333	74	413	162
651 MHSA: Anxiety disorders per 100,000	9101	4450	9101	4486	8780	3877	8065	2957	8212	3454	9222	4630	9210	5018
652 MHSA: Attention-deficit, conduct, and disruptive behavior disorders per 100,000	667	299	667	260	547	247	672	0	384	384	481	222	731	389
653 MHSA: Delirium, dementia, and amnestic and other cognitive disorders per 100,000	604	2914	604	2905	677	2837	1075	5108	460	1305	370	1667	659	3766
654 MHSA: Developmental disorders per 100,000	349	394	349	345	417	387	538	403	384	384	259	370	353	455
655 MHSA: Disorders usually diagnosed in infancy, childhood, or adolescence per 100,000	49	92	49	103	87	73	269	269	0	153	37	111	48	96
656 MHSA: Impulse control disorders, NEC per 100,000	31	17	31	27	47	10	134	0	0	77	111	0	18	18
657 MHSA: Mood disorders per 100,000	20459	6050	20459	6039	20257	5053	22446	3360	20568	5219	16704	3926	21024	7653
658 MHSA: Personality disorders per 100,000	1079	117	1079	112	1437	80	403	0	3761	77	593	74	1000	156
659 MHSA: Schizophrenia and other psychotic disorders per 100,000	5489	1705	5489	1589	7297	1757	7796	1344	10821	2609	5370	1370	4946	1683

660 MHSA: Alcohol-related disorders per 100,000	13166	3562	13166	3599	9743	3027	6048	672	9670	3607	9222	3815	14485	4210
661 MHSA: Substance-related disorders per 100,000	9607	2981	9607	2771	9597	2603	4973	1075	14121	4068	9630	2407	9425	3437
662 MHSA: Suicide and intentional self-inflicted injury per 100,000	4866	647	4866	591	5117	560	4167	269	7675	691	4407	556	4814	772
663 MHSA: Screening and history of mental health and substance abuse codes per 100,000	19698	12274	19698	11994	15433	10963	11425	9140	15272	15196	12778	8074	21766	15623
670 MHSA: Miscellaneous mental health disorders per 100,000	792	395	792	381	730	410	1478	403	153	537	667	519	856	371

Model Descriptions

Logistic regression assumes a linear relationship between features and the log odds of suicide death. This method has no hyperparameters; training via maximum likelihood estimation is fast. However, the linear relationship may not be true, leading to less accurate predictions. Because it is widely used, we include logistic regression as a baseline. Naive Bayes is a probabilistic classifier that assumes conditional independence of the features given the class. We include it because it is scalable to large data sets, can handle particular types of dependencies between features, and has proven useful for real-world problems including suicide prediction[2,3]. The remaining two methods we employ, random forests and gradient boosted trees, build nonlinear models using ensembles of decision trees[4,5]. To train gradient boosted trees and random forests, we use XGBoost[6]; all other models are trained using scikit-learn[7].

When modeling with random forests we use 100 trees; we find that the optimal tree depth depends on the resampling method chosen. Increasing depth increases specificity at the expense of sensitivity; essentially, deeper trees overfit the negative class. For gradient boosted trees, we achieve the best results with trees of depth at most 2, but generally a depth of 1 was preferable. For choosing the depth parameter for random forests and gradient boosted trees we chose one depth and used that parameter for each of the 10 runs, we did not select a new depth parameter for each of the reshuffled data sets.

For each of the 10 runs for each model type we shuffle the data with a new random seed by patient RLN and split the data into training/validation/test sets by RLN. This ensures the patient groups are disjoint across the training, validation, and test sets. We then run each model with parameters chosen from a single run.

For all models, we adjust the threshold τ so that training specificity equals 0.76. This approximately equalizes test set specificities, enabling comparison of models based on test set sensitivity. Like the depth parameter, we selected a τ parameter based on one run of the model and applied this value for each of the 10 runs.

To calculate average model range, we calculated the range for each run of the model and then report the average and the standard deviation. We do not report the range of the averaged values by race.

We also present the quantile-quantile plot of the results of 100 logistic regression models for sensitivity Figure S1, specificity Figure S2, and AUC Figure S3. This shows that the results from the models with different random seeds are normally distributed and that the standard deviation metric is informative.

Due to the size of the predictor space and the model, printing the complete model is not feasible nor do we find that it would be very informative. However, upon request, we are happy to share the trained models as saved Python objects, together with instructions on how to apply these trained models, e.g., with synthetic data, to generate predictions.

Results

Fairness Metrics

We begin with the following premise: we seek models for which *the opportunity to receive treatment is independent of racial/ethnic group identity, conditional on the true outcome*. In this regime, all racial/ethnic groups would have equal opportunities to receive outreach services and interventions to prevent suicide death.

We formalize this in the language of probability by defining three random variables: A, $\hat{Y}$, and Y. Here A denotes racial/ethnic identity (or, more generally, any protected attribute), $\hat{Y}$ denotes our model's prediction, and Y denotes the true label[8]. Then, we can express our goal (the italicized clause above) as follows:

$$\Pr\{\hat{Y} = 1 \,|\, A = race, Y = y\} = \Pr\{\hat{Y} = 1 \,|\, Y = y\} \quad y \in \{0,1\}$$

In words, this equation states that regardless of racial/ethnic group identity, each patient will have the same probability of receiving treatment.

Following the laws of probability, the above equation implies the *equal odds*[8] criterion

$$\Pr\{\hat{Y} = 1 \mid A = race1, Y = y\} = \Pr\{\hat{Y} = 1 \mid A = race2, Y = y\} \quad y \in \{0,1\}$$

If we only enforce this criterion on records that correspond to patients who have died by suicide, we obtain the *equal opportunity*[8] criterion

$$\Pr\{\hat{Y} = 1 \mid A = race1, Y = 1\} = \Pr\{\hat{Y} = 1 \mid A = race2, Y = 1\}$$

In our context, equal opportunity is equivalent to balancing the model's sensitivity or TPR (true positive rate) across racial/ethnic groups. Equal odds is equivalent to balancing both the TPR and the FPR (false positive rate) across racial/ethnic groups. As specificity is 1-FPR, we see that equal odds is equivalent to balancing both sensitivity and specificity across racial/ethnic groups.

In Table 2 (sensitivity), a range of zero indicates that conditional on a patient's predictors, the probability of $\hat{Y} = 1$ (suicide death) would not depend on the patient's racial/ethnic group (i.e., the equal opportunity criteria)[8]. If the same model also had zero range in Table 3 (specificity), that would indicate that conditional on a patient's predictors, the probability of either $\hat{Y} = 1$ or $\hat{Y} = 0$ would not depend on the patient's racial/ethnic group (i.e., the equal odds criteria)[8].

In addition to the Separate and Equity models fulfilling equal odds criteria, we also see a major flaw with the Blind method. When training with the Blind method, White patient files (a majority of the data set) are much more likely to be classified as positive for suicide death regardless of the true label (high sensitivity, low specificity). The opposite is true for all minority groups with patient files much less likely to be positive for suicide death regardless of true label (low sensitivity, high specificity). This model's overreliance on race/ethnicity features dominates whatever it learns about other features that predict suicide death.

Additional Results

In the main text, we report average model performance with standard deviation on data that has been shuffled, split and resampled with 10 random seeds to ensure the robustness of these methods. To show that reporting the average with standard deviation is a meaningful metric we run 100 simulations on the logistic regression model to show the sensitivity, specificity, and AUC metrics are normally distributed. This is shown in Figure S1-S3.

In this work, we only report performance on the test set. We do not report performance on developmental data (train and validation set) because (i) metrics on the training set would be slightly inflated due to the fact the model learns from the training set and (ii) the validation set results would be very similar to the test set results.

Figure S1 Quantile-quantile plot for sensitivity of 100 runs of a logistic regression model. This shows the performance on the sensitivity metric is normally distributed over 100 runs. Figure produce by author (MR).

Figure S2 Quantile-quantile plot for specificity of 100 runs of a logistic regression model. This shows the performance on the specificity metric is normally distributed over 100 runs. Figure produce by author (MR).

Figure S3 Quantile-quantile plot for AUC of 100 runs of a logistic regression model. This shows the performance on the AUC metric is normally distributed over 100 runs. Figure produce by author (MR).

References

1 Collins GS, Reitsma JB, Altman DG, *et al.* Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *European Urology* 2015;**67**. doi:10.1016/j.eururo.2014.11.025

2 Zhang H. The Optimality of Naive Bayes. In: Barr V, Markov Z, eds. *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, Miami Beach, Florida, USA*. AAAI Press 2004. 562–7.http://www.aaai.org/Library/FLAIRS/2004/flairs04-097.php

3 Barak-Corren Y, Castro VM, Javitt S, *et al.* Predicting suicidal behavior from longitudinal electronic health records. *American Journal of Psychiatry* 2017;**174**:154–62.

4 Hastie T, Tibshirani R, Friedman JH. *The Elements of Statistical Learning*. Second. New York, NY, USA: : Springer 2009.

5 Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Rrees*. 2017. doi:10.1201/9781315139470

6 Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. 785–94.

7 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011;**12**:2825–30.

8 Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems*. 2016. 3315–23.