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.