Journal of the American Medical Informatics Association, 28(11), 2021, 2393–2403
doi: 10.1093/jamia/ocab148
Advance Access Publication Date: 12 August 2021
Research and Applications

Research and Applications

Bias and fairness assessment of a natural language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups

Hale M. Thompson,[1] Brihat Sharma,[1] Sameer Bhalla,[1] Randy Boley,[1] Connor McCluskey,[1] Dmitriy Dligach,[2] Matthew M. Churpek,[3] Niranjan S. Karnik,[1] and Majid Afshar[3]

[1]Department of Psychiatry & Behavioral Sciences, Rush University Medical Center, Chicago, Illinois, USA, [2]Department of Computer Science, Loyola University, Chicago, Illinois, USA, and [3]Department of Medicine, University of Wisconsin, Madison, Wisconsin, USA

Corresponding Author: Hale M. Thompson, PhD, Department of Psychiatry & Behavioral Sciences, Rush University Medical Center, 1645 W. Jackson Blvd., Suite 302, Chicago, IL 60612, USA (hale_thompson@rush.edu)

Received 18 April 2021; Revised 28 June 2021; Editorial Decision 29 June 2021; Accepted 1 July 2021

ABSTRACT

Objectives: To assess fairness and bias of a previously validated machine learning opioid misuse classifier.
Materials & Methods: Two experiments were conducted with the classifier's original (n = 1000) and external validation (n = 53 974) datasets from 2 health systems. Bias was assessed via testing for differences in type II error rates across racial/ethnic subgroups (Black, Hispanic/Latinx, White, Other) using bootstrapped 95% confidence intervals. A local surrogate model was estimated to interpret the classifier's predictions by race and averaged globally from the datasets. Subgroup analyses and post-hoc recalibrations were conducted to attempt to mitigate biased metrics.
Results: We identified bias in the false negative rate (FNR = 0.32) of the Black subgroup compared to the FNR (0.17) of the White subgroup. Top features included "heroin" and "substance abuse" across subgroups. Post-hoc recalibrations eliminated bias in FNR with minimal changes in other subgroup error metrics. The Black FNR subgroup had higher risk scores for readmission and mortality than the White FNR subgroup, and a higher mortality risk score than the Black true positive subgroup ($P < .05$).
Discussion: The Black FNR subgroup had the greatest severity of disease and risk for poor outcomes. Similar features were present between subgroups for predicting opioid misuse, but inequities were present. Post-hoc mitigation techniques mitigated bias in type II error rate without creating substantial type I error rates. From model design through deployment, bias and data disadvantages should be systematically addressed.
Conclusion: Standardized, transparent bias assessments are needed to improve trustworthiness in clinical machine learning models.

Key words: structural racism, bias and fairness, machine learning, natural language processing, opioid use disorder, interpretability

INTRODUCTION

Intelligent computing and machine learning have gained prominence for their roles in patient-centered healthcare and the democratization of medicine over the last 20 years.[1–3] Machine learning classifiers have been shown to outperform clinical judgment at the population level by reducing screening burdens and access inequities for common illnesses and conditions.[4–6] As with any data, analytical techniques brought to them can contain a range of biases, from sample bias to measurement bias to representation bias and historical bias. In machine learning, bias impacts every step of model development, testing, and implementation and can lead to algorithmic bias and feedback loops known as biased network effects.[7] Biases tend to unfairly disadvantage some groups or populations over others—often those already disproportionately marginalized.[8,9] When natural language processing (NLP) classifiers are developed with biased or imbalanced datasets,[10] disparities across subgroups may be codified, perpetuated, and exacerbated if biases are not assessed, identified, mitigated, or eliminated. In healthcare settings, these biases and disparities can create multiple layers of harm.[9,11]

In the 21st century, US medical institutions and pharmaceutical companies have been a key driver of the first of a triple-wave opioid overdose epidemic.[12] Opioid prescribing tripled between the 1990s and 2011, and opioid overdose deaths due to pharmaceutical opioids more than tripled during that time and was tied to an older, whiter age cohort than that of the late 20th century.[13] With the distribution and consumption of opioids shifting to pharmaceuticals and patients with pain, opioid misuse treatment expanded from the criminalization and abstinence models that had mainly targeted urban Black and Brown men, toward a disease or addiction model; this shift has been associated with a White middle-class logic of eugenics and neuroscience[14,15] that has made space to sympathize with both White rural and suburban opioid consumers and the corporate pharmaceutical companies and doctors that distribute them.[16] Studies of universal substance misuse screening programs[17] and treatment services[18] show how medicine continues to codify and perpetuate racial biases and access inequities.

OBJECTIVES

Given the structural and historical backdrops that impact clinical data regarding substance misuse, we are operationalizing principles of fairness, accountability, transparency, and ethics (FATE) to assess our NLP opioid misuse classifier's predictions.[11,19] The identification of bias and fairness in screening tools is critical to plan for mitigation or elimination prior to deployment. In this article, we first apply techniques to audit our classifier's fairness and bias by adapting a bias toolkit[20] and then attempt to correct bias with post-hoc methods. Second, we examine face validity by running Local Interpretable Model-Agnostic Explanations (LIME)[21] across all individual predictions and providing averaged features to further assess for differences in features between race/ethnic groups. We believe this study is a key step toward a more transparent assessment of machine learning models .

MATERIALS AND METHODS

Development and internal validation dataset

The opioid misuse classifier was originally developed from hospitalized patient data at Loyola University Medical Center (Loyola). Loyola is a 559-bed hospital and tertiary academic center, including a burn and Level 1 trauma center, serving Chicago and its western suburbs. The study cohort for annotation consisted of a sampling (n = 1000) of adult hospital encounters from the electronic health record (EHR) between 2007 and 2017. Oversampling was performed for hospitalizations with International Classification of Diseases (ICD)-9 and 10 codes related to opioid misuse or chronic pain, urine drug screens positive for opioids, naloxone orders and administration (ie, for opioid overdoses), or physician orders for urine drug screens (signifying at-risk individuals). The final dataset for development and internal validation consisted of 367 manually labeled cases, age- and sex-matched with controls that had no indications of opioid misuse.

The final classifier was a standardized vocabulary embedding into a Convolutional Neural Network (CNN), and it outperformed a rule-based classifier developed by addiction experts as well as other machine learning classifiers. The previously developed CNN opioid classifier is accessible at https://github.com/AfsharJoyceInfoLab/OpioidNLP_Classifier. The CNN opioid classifier demonstrated a sensitivity of 79% (95% CI: 68%–88%) and a specificity of 91% (95% CI: 85%–95%). For additional details on model development see Sharma and colleagues.[4]

External validation dataset

The external validation dataset was derived from the EHR at Rush University Medical Center (Rush). Rush is a 727-bed hospital, tertiary care academic center located on the West Side of Chicago. Rush launched a multidisciplinary Substance Use Intervention Team (SUIT) to address the opioid epidemic through a Screening, Brief Intervention, and Referral to Treatment (SBIRT) program with an inpatient Addiction Consult Service in October 2017.[22,23] Part of the SUIT initiative included the following single question universal drug screen: "How many times in the past year have you used an illegal drug or used a prescription medication for non-medical reasons?" ($\geq$1 is positive). The single-question screen was administered by nursing staff as a part of the admission battery of questions to patients admitted to Rush's 18 inpatient medical and surgical wards. Nursing staff were encouraged to complete the question but, as with much of the admission battery, a forced response was not required. Patients with a positive universal screen were referred for a full screen with the 10-item Drug Abuse Screening Test (DAST-10).[24] Survey data collected during the hospital-wide screening program served as the reference dataset for external validation of the opioid classifier that was developed at Loyola. The inclusion criteria were all unplanned adult inpatient encounters ($\geq$18 years of age) who were screened between October 23, 2017, and December 31, 2019 (n = 53 974).

In external validation using the first 24 hours of clinical notes, the CNN opioid classifier demonstrated a sensitivity of 80% (95% CI: 77%–83%) and a specificity of 99% (95% CI: 99%–99%). Screened patients with opioid misuse were disproportionately younger, male, and Black compared with the patients with no misuse, and also on Medicaid and discharged against medical advice compared to those with no misuse.[25] For a more detailed description of the external validation methods, see Afshar and colleagues' publication.[25]

Analysis plan

To conduct our experiments for the assessment of bias and fairness, we used the external validation data (n = 53 974) as representative of patients that participated in a hospital-wide screening program.

Given racial disparities around opioid misuse, the demographics of our patient population, and the attention to systemic racism in the US in 2020, we prioritized auditing the models for parity by race. In Chicago, the heroin and overdose epidemics have had the greatest impact on non-Hispanic Black persons.[26] Of note, the racial categories in the Rush EHR system are White, Black, Asian, American Indian or Alaskan Native, Native Hawaiian or Other Pacific Islander, Other, and Missing/Declined. Ethnicity is collected as Hispanic/Latinx or non-Hispanic/Latinx. As more Hispanic/Latinx patients choose "Other" for race (n = 6307), we adopted CDC methods and those of large, national health surveys and combined "Hispanic/Latinx—Other Race" with Hispanic/Latinx White and Hispanic/Latinx Black ethnicity (n = 3089) to make "Hispanic/Latinx" race/ethnicity (n = 9252).[27–29] The 4 remaining racial/ethnic categories—Asian, Native American or Alaskan Native, Native Hawaiian or Other Pacific Islander, Other Race/Ethnicity, plus refuse/unknown—have been collapsed into Other (n = 3836). In the following analysis, non-Hispanic/Latinx White and non-Hispanic/Latinx Black are referred to as White and Black, respectively. We conducted 2 distinct experiments to assess bias of the classifier's opioid misuse predictions. We included the 4 racial/ethnic patient subgroups from the external validation dataset: White (n = 23 345), Black (n = 17 541), Hispanic/Latinx (n = 9252), and Other (n = 3836).

In the first experiment, we adapted a publicly available python toolkit for auditing ML models for fairness and bias, and used the same group distributional and error-based metrics.[20] As an intervention, our opioid misuse classifier is assistive rather than punitive. In other words, the goal of the classifier is to provide point-of-care education, treatment options, and care pathways to patients who misuse opioids. Therefore, we tested for disparities in type II errors (ie, false negative classifications) across groups, stratified by age range, sex, and race/ethnicity, since we do not want to miss treatment opportunities due to such disparities. To assess bias and fairness, we measured each subgroup's false negative rate (FNR), which is the fraction of false negatives of a subgroup among the number of labeled positives. We did not prioritize the false omission rate (FOR) which was too sparse to measure; this dataset is highly imbalanced with approximately 99% of encounters labeled negative. Focusing on race/ethnicity, our model was considered biased if the FNR of Black, Hispanic/Latinx, or Other is greater than the White FNR, and if the predicted positive rate (PPR) of Black, Hispanic/Latinx, or Other vs White does not reflect statistical parity. The PPR is the fraction of cases predicted as positive within a group and is a distributional metric and not a measure of type II error. Although PPR has less impact at the patient-level than FNR, particularly for an assistive medical intervention like substance misuse prevention and treatment, we aimed for statistical parity of PPR between subgroups. Comparison between subgroups, including race/ethnic groups, were made using bootstrapped 95% confidence intervals (CI) to detect disparities in the group distributional and error-based point estimates.

Our second experiment used Local Interpretable Model-Agnostic Explanations (LIME)[30] to estimate the face validity, also referred to as local fidelity,[31] of the opioid misuse classifier's predictions on the Rush dataset. LIME trains an interpretable model by generating local surrogate models to explain the individual predictions of the CNN, opioid misuse classifier. Rather than try to test the entirety of features generated by the opioid misuse classifier, our surrogate model made individual or local predictions of the external validation data by racial subgroup and compared them to the global predictions from the training data from Loyola. In other words, LIME estimated the global interpretability of the CNN model in order to examine its face validity. For the CNN model, a local surrogate model was applied to approximate the predictions to explain individual predictions locally and then average the feature weights from the local explanations to derive a global measure across all patients. The global LIME measure had an average median R^2 (variance explained) of 0.979 (IQR 0.972–0.984), which was an excellent approximation for the CNN opioid misuse classifier. LIME results were examined across subgroups to assess for any differences in the features between race/ethnic subgroups that could be removed for presumed bias.

Two post-hoc bias mitigation experiments were conducted: the first by varying the cut point in the subgroup with the biased FNR and the other by recalibrating the classifier by subgroup. The cut point was varied in the subgroup with a biased FNR estimate to improve sensitivity without losing specificity. In our prior publications, we examined a range of cut points, including the Youden index,[32] to identify the optimal sensitivity and specificity for a hospital-wide program. The same approach was followed during subgroup recalibration. Model predictions for the external validation dataset were recalibrated using isotonic calibration[33] to better match the observed events with the predicted probabilities by the opioid misuse classifier. After each approach, we reran our bias assessment to examine for mitigation.

To examine the discordance between the ground truth labels and predicted labels, we conducted a chart review of all the false negatives across the Black subgroup from our external validation dataset. Two trained reviewers (SB and CM) with high interrater agreement (Kappa score = 0.90) performed chart reviews to verify the level of opioid misuse using previously developed guidelines.[25] Patient characteristics across subgroups and by misclassification type were analyzed. The Elixhauser classification of 30 diagnostic codes, accounting for major comorbidities including mental health conditions and substance use disorder diagnoses, was the basis for mortality and readmission scores.[34] We used chi-squared tests for categorical variables and the Kruskal-Wallis H-test for continuous variables ($P < .05$). Statistical analyses were conducted with Python Version 3.6.5 (Python Software Foundation).

The Rush University (#18061108) and Loyola University Chicago (LU #209950) Institutional Review Boards approved these analyses and waived informed consent for use of retrospective patient data.

RESULTS

In the external validation dataset, the true negative rate from our opioid misuse classifier was consistent across all subgroups, including race. However, the classifier had a decrease in the true positive rate with increasing age groups (Table 1). We also identified bias in the false negative rate (FNR) of the Black subgroup (n = 106) when compared to the White subgroup (n = 34). The opioid misuse classifier's FNR was higher among the Black subgroup (0.32; 95% CI: 0.27–0.37) in comparison to the White subgroup (0.17; 95% CI: 0.12–0.23) (Figure 1). The predicted positive rate (PPR) was also higher among the Black subgroup (0.51; 95% CI: 0.48–0.55) compared to the White subgroup (0.33; 95% CI: 0.30–0.37).

The LIME experiment demonstrated good face validity among the top features of the original and external validation datasets; however, no differences were noted between subgroups. "Heroin" was the most important global feature across all positive cases (Figure 2). In the Black and White subgroups, the top features were

Table 1. Test characteristics and 95% confidence intervals across age, sex, and racial/ethnic subgroups of the external validation cohort (N = 53 974)*

	True positive rate	True negative rate	False positive rate	False negative rate	Precision
Age in years					
18–44	0.837 (0.787–0.880)	0.991 (0.990–0.993)	0.009 (0.007–0.011)	0.163 (0.121–0.213)	0.646 (0.593–0.697)
45–60	0.731 (0.672–0.784)	0.991 (0.989–0.993)	0.009 (0.008–0.011)	0.270 (0.216–0.328)	0.599 (0.543–0.654)
61–70	0.588 (0.483–0.687)	0.996 (0.995–0.997)	0.004 (0.003–0.005)	0.412 (0.313–0.517)	0.533 (0.434–0.630)
≥ 71	0.273 (0.060–0.610)	1.000 (0.999–1.0)	0.0003 —	0.727 (0.390–0.940)	0.429 (0.099–0.816)
Sex					
Female	0.776 (0.717–0.829)	0.996 (0.995–0.996)	0.005 (0.004–0.005)	0.224 (0.171–0.283)	0.557 (0.500–0.612)
Male	0.728 (0.681–0.770)	0.993 (0.992–0.994)	0.007 (0.006–0.008)	0.273 (0.229–0.319)	0.647 (0.573–0.695)
Race/ethnicity					
Black	*0.685* *(0.632–0.734)*	0.991 (0.989–0.992)	0.010 (0.008–0.011)	*0.316* *(0.266–0.368)*	0.585 (0.535–0.634)
Hispanic/Latinx	0.833 (0.727–0.911)	0.997 (0.996–0.998)	0.003 (0.002–0.004)	0.167 (0.089–0.273)	0.698 (0.589–0.792)
White	0.827 (0.766–0.877)	0.996 (0.995–0.999)	0.004 (0.003–0.005)	0.174 (0.123–0.234)	0.635 (0.573–0.695)
Other (0.447–0.844)	0.667 (0.993–0.997)	0.995 (0.003–0.008)	0.005 (0.153–0.553)	0.333 (0.298–0.649)	0.471

*Significant point estimates and confidence intervals for race have been italicized.

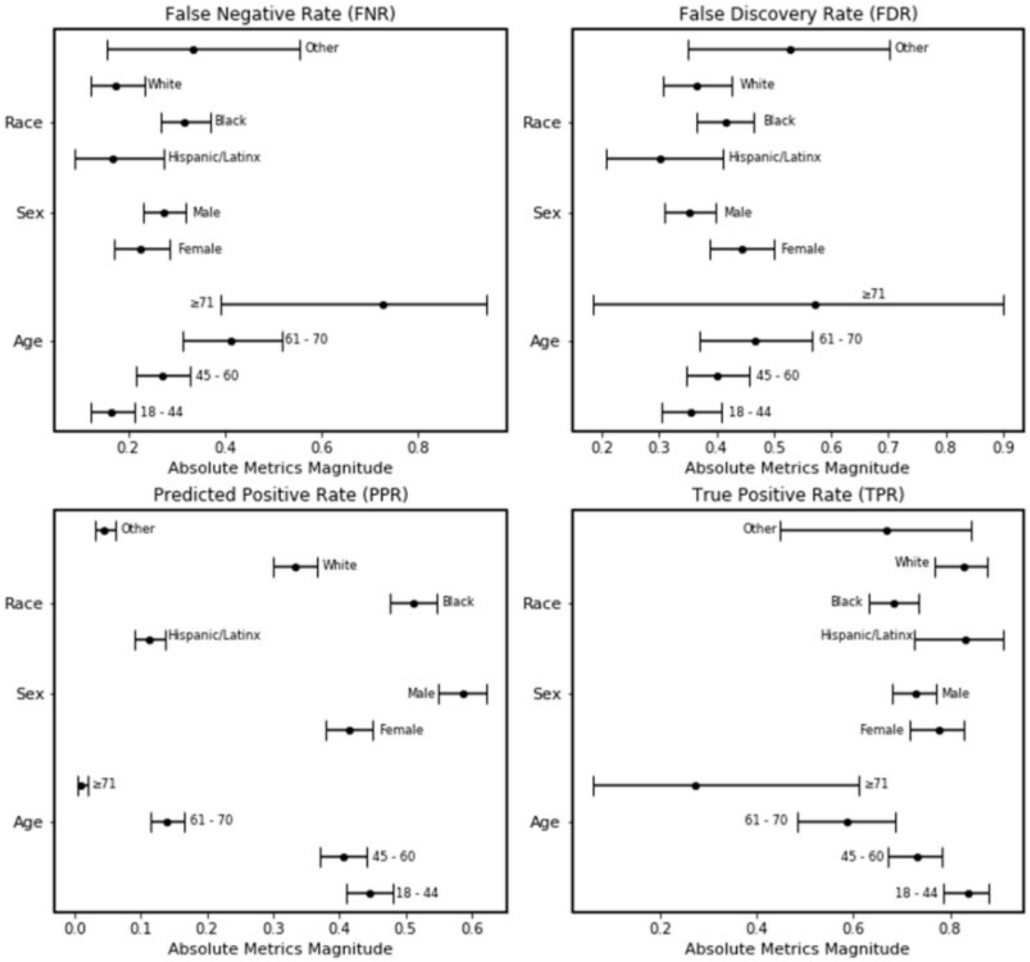

Figure 1. Plot of bias and fairness point estimates with bootstrapped 95% confidence intervals for the NLP opioid misuse classifier's predictions for the external validation cohort.

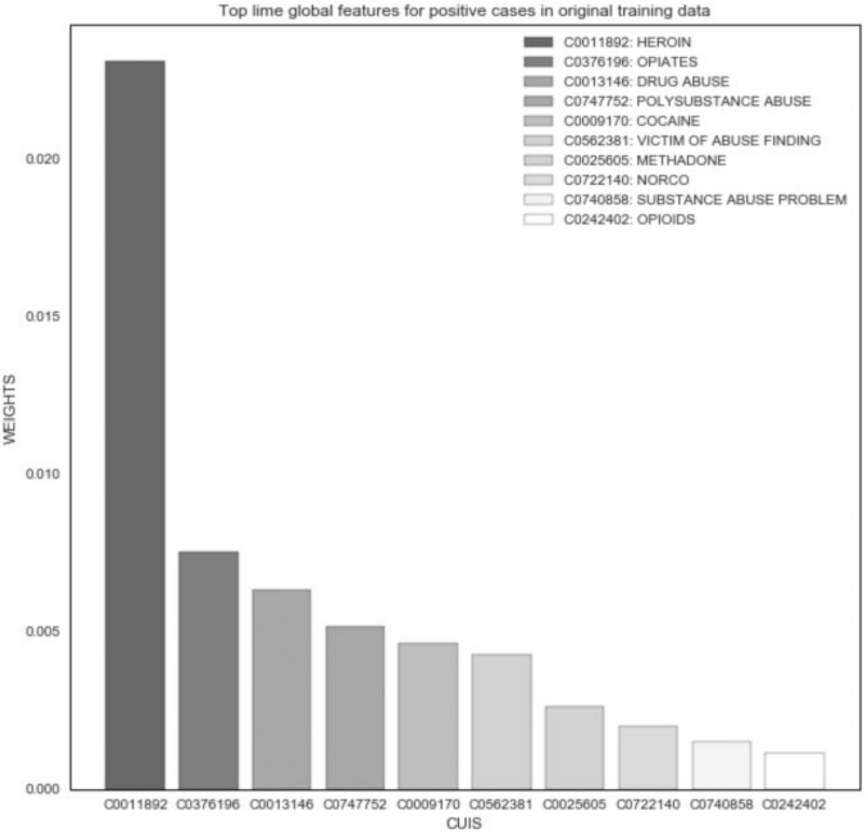

Figure 2. NLP opioid misuse classifier's top features for positive cases in original development dataset (2007-2017).

very similar; "heroin" was the top feature for Black positive cases (Figure 3), and "drug abuse" was the top feature for White (Figure 4). Only the weights differed in comparison to global features of the original dataset. For example, the average weight of "heroin" in the original dataset is nearly twice that of the Black subgroup (0.014) and 10 times greater than that of the White subgroup (0.002) in the external validation dataset.

Our first mitigation experiment reduced the cut point in the Black subgroup from 0.3 to 0.2; this change improved the sensitivity in the subgroup from 0.68 to 0.75. The resulting FNR for the Black subgroup was reduced to 0.25 (95% CI:0.20–0.30) and in closer approximation to the White subgroup with overlapping confidence intervals (Figure 5). However, the FDR for the Black subgroup increased from 0.41 (95% CI: 0.37–0.47) to 0.46 (95% CI: 0.41–0.50) but still overlapped with the White subgroup's FDR (0.36; 95% CI: 0.31–0.43). The disparity in PPR between Black and White subgroups widened; the Black subgroup PPR increased from 0.51 (95% CI: 0.46–0.55) to 0.55 (95% CI: 0.52–0.59) and the White subgroup decreased from 0.33 (95% CI: 0.30–0.37) to 0.30 (95% CI: 0.27–0.34).

The second mitigation technique recalibrated the classifier by subgroup. Again, the recalibration removed the bias in the FNR of the Black subgroup (Figure 6). The recalibrated FNR for the Black subgroup was reduced to 0.24 (95% CI: 0.19–0.29) relative to the White subgroup FNR of 0.21 (95% CI: 0.15–0.27). The FDR for the Black subgroup increased from 0.41 (95% CI: 0.37–0.47) to 0.46 (95% CI: 0.41–0.50) but still overlapped with the White subgroup's FDR (0.36; 95% CI: 0.31–0.43). The disparity in PPR between Black and White subgroups widened; the Black subgroup PPR increased from 0.51 (95% CI: 0.46–0.55) to 0.56 (95% CI: 0.53–0.60) and the White subgroup decreased from 0.33 (95% CI: 0.30–0.37) to 0.29 (95% CI: 0.26–0.32).

Post-hoc chart review confirmed that 98% of the false negative encounters were, in fact, false negatives. Compared to the FNR of the White subgroup, the Black FNR subgroup is older (58 vs 42 years, $P < .01$) and has a higher median readmission score (47 vs 36, $P < .01$) and a higher median mortality score (7 vs −1, $P < .01$) (Table 2). The Black subgroup had greater proportions in 6 Elixhauser comorbidities ($P < .05$): complicated hypertension, obesity, renal failure, congestive heart failure, chronic lung disease, and drug-related diagnoses. The Black and White subgroups had similar distributions of sex and insurance type, whereby males were 72% among Black subgroup encounters and 65% among the White ($P = .44$), and Medicaid represents 70% (Black) and 68% (White) of each subgroup's encounters ($P = .14$).

Between the false negative rate (FNR) and true positive rate (TPR) for the Black subgroup, there were no differences across distributions of age, sex, type of insurance, or the median readmission score (Table 3). The main difference was the median mortality score (7 vs 3, $P < .01$). The Black FNR subgroup had greater proportions across 8 disease comorbidities, for example complicated hypertension (50% vs 9%, $P < .01$), and had less drug-related (94% vs 99%, $P = .02$), psychiatric (8% vs 20%, $P < .01$), and AIDS comorbidities (0 vs 7%, $P < .01$), compared to the Black TPR group.

DISCUSSION

To our knowledge, this is the first assessment of interpretability, bias, and fairness of an NLP classifier for substance misuse. Our assessment

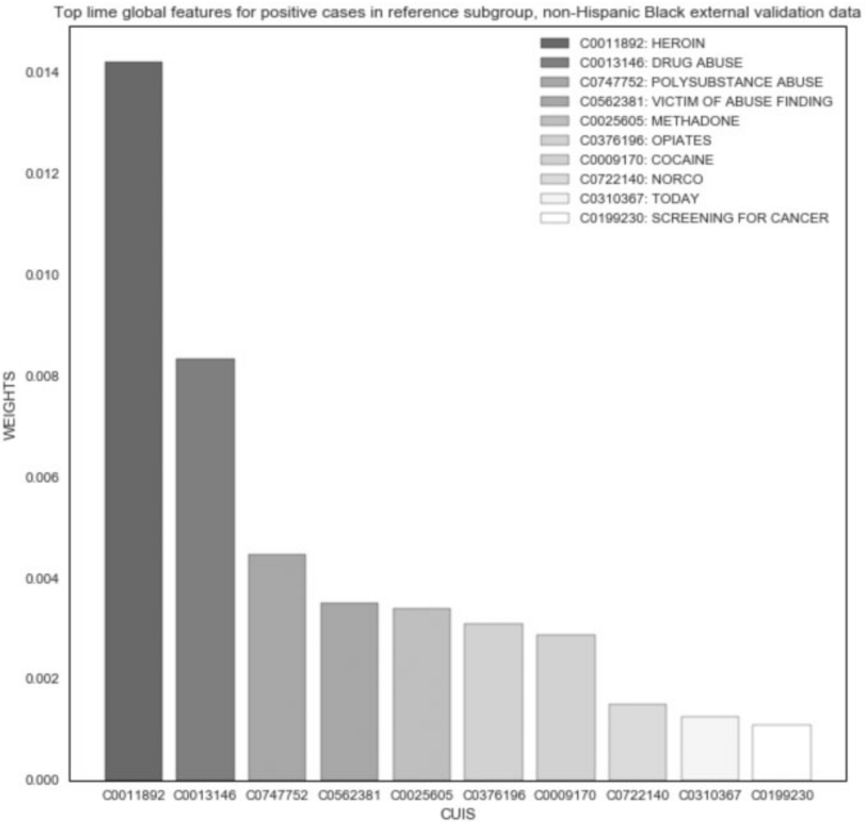

Figure 3. NLP opioid misuse classifier's top features for positive cases in Black subgroup of external validation dataset (2017–2019).

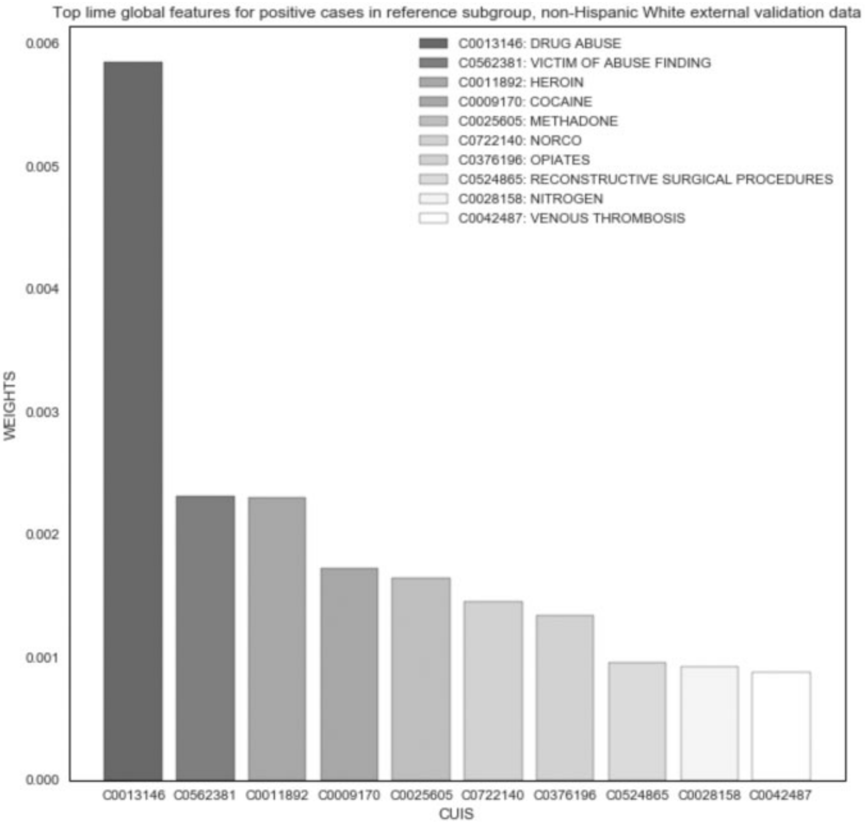

Figure 4. NLP opioid misuse classifier's top features for positive cases in White subgroup of external validation dataset (2017–2019).

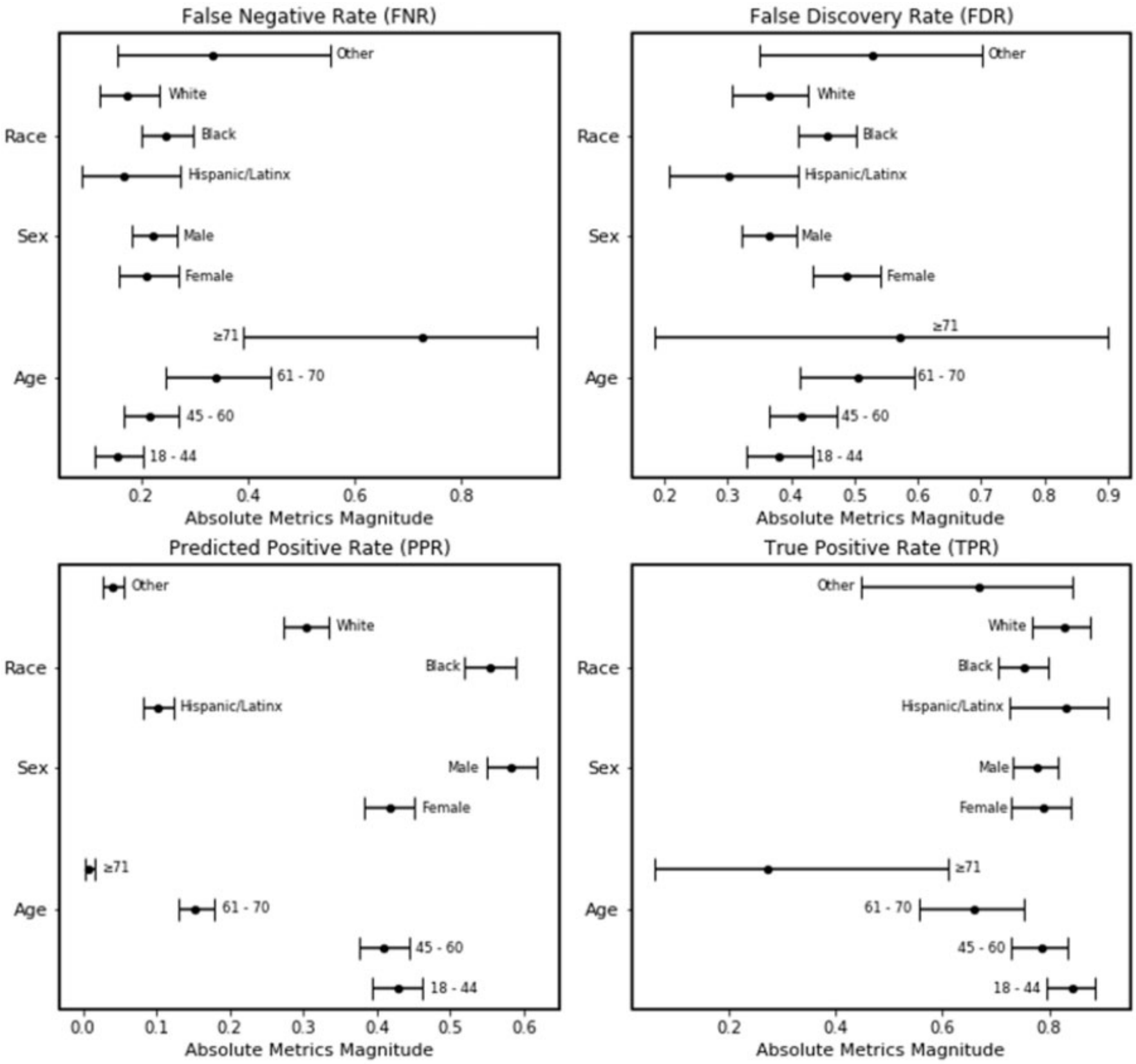

Figure 5. Plot of bias and fairness metrics of the NLP opioid misuse classifier's prediction for the external validation cohort with cut point adjustment by subgroup.

of the opioid misuse classifier identified bias and unfairness in the Black FNR subgroup compared to the White FNR subgroup. We demonstrate some success in our post-hoc mitigation experiments to mitigate bias in the false-negative rate for the more disparate Black subgroup. The LIME experiment confirmed that the top features are similar in both substance and directionality between the training–testing dataset and the external validation dataset by subgroup. The estimated weights of those top features in the validation dataset, however, are weaker in the subgroups. This difference between the feature weights is likely due to the relatively large imbalance of the Rush external validation dataset where 628 encounters were predicted positive out of 53 974 compared to 367 positive cases out of 1000 in the original dataset. The analysis of the FNR by Black and White patient subgroups' data indicate that the Black patients are older and sicker, requiring more medically complex treatments for comorbidities than the White FNR subgroup. Similarly, the Black FNR subgroup has higher proportions of comorbidities and a higher mortality score compared to the Black subgroup of true positives. These data disadvantages biasing the classifier's predictions may be tied to the comorbidity and mortality disparities; notably, these disparities are consistent with relatively poor health outcomes among the adult Black population in the United States and may reflect deeper structural inequities around access to primary care and trust in medical institutions.[35–37]

Our experiments demonstrate the importance of transparency in machine learning model development and validation as a critical step toward building the public's trust in applications of artificial intelligence in medicine. Because the SUIT intervention is assistive, we are concerned about missing the opportunity to treat Black patients for opioid misuse; so the FNR was a focus in our mitigation efforts. Although the disparity in predicted positive rate grew with our bias mitigation efforts for the FNR, the increase in PPR of the Black patient subgroup is likely due to the increase in true positives. Also, the relatively small increase in the FDR after mitigating the FNR is worth noting and may translate into more harm by driving overtreatment or misdiagnosis of Black patients for opioid misuse—increasing both stigma for Black patients and their mistrust of providers—while also increasing "alarm fatigue" for providers.

Other studies have taken different approaches to mitigate bias in clinical machine learning models, including a bias correction method based on learning curve fitting and removing the features that exac-

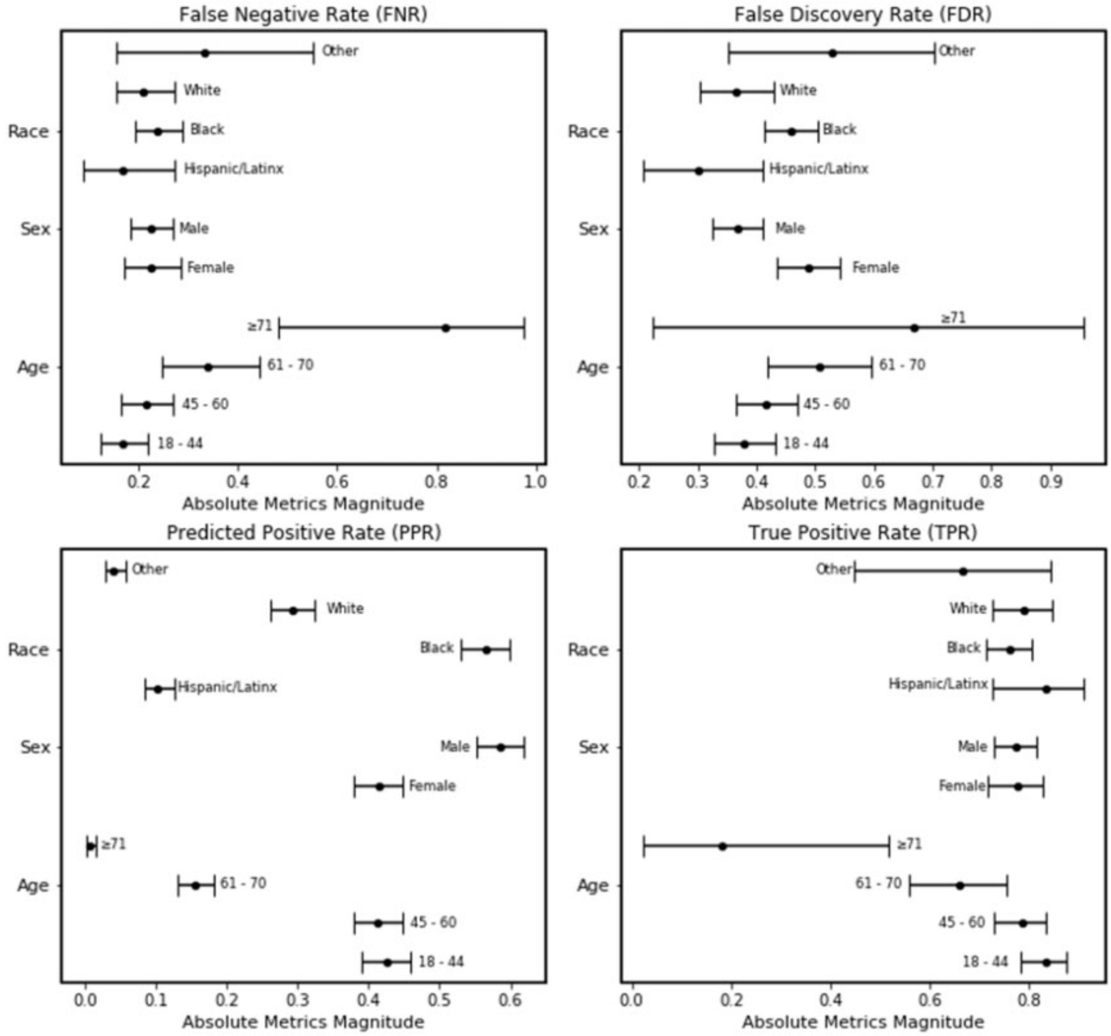

Figure 6. Plot of bias and fairness metrics of the NLP opioid misuse classifier's prediction for the external validation cohort after model recalibration by subgroup.

Table 2. Opioid classifier external validation patient characteristics of false negative predictions for opioid misuse, comparing Black and White subgroups (2017–2019)

	FNR Black subgroup		FNR White subgroup		df	*P* value
Age						
Median (IQR)	*58*	*(52.25–64.75)*	*42*	*(32.5–48.75)*		*<.01*
Sex	n = 106	%	n = 34	%		
Female	30	0.28	12	0.35	1	.44
Male	76	0.72	22	0.65		
Insurance						
Medicaid	74	0.7	23	0.68	2	.14
Medicare	21	0.2	6	0.18		
Private	11	0.1	5	0.15		
Readmission score						
Median (IQR)	*47*	*(33–62.5)*	*35.5*	*(23–47.75)*		*<.01*
Mortality score						
Median (IQR)	*7*	*(−1–19)*	*−1*	*(−5–13)*		*<.01*

Table 3. Opioid misuse classifier external validation patient characteristics of Black subgroup comparing false negative to true positive for opioid misuse (2017–2019)

Black subgroup						
	False negative		True positive			
	n = 106		n = 230		df	*P* value
Age						
Median (IQR)	*58*	*(52.25–64.75)*	*55*	*(49–60)*		*<.01*
Sex	**n**	%	**n**	%		
Female	30	0.28	84	0.37	1	.14
Male	76	0.72	146	0.63		
Insurance						
Medicaid	74	0.7	167	0.73	2	.33
Medicare	21	0.2	36	0.16		
Private	11	0.1	27	0.12		
Readmission score						
Median (IQR)	47	(33–62.5)	43	(30.25–57)		.09
Mortality score						
Median (IQR)	*7*	*(−1–19)*	*3*	*(−6–12)*		*<.01*

erbate bias.[38,39] However, in our case, we worked with a pragmatic dataset that had a fixed sample size, and our LIME experiments did not reveal major disparities in CUIs between racial subgroups. Our recalibration approach is an average approach across all predicted probabilities but methods in decision curve analysis may be applied to better delineate differing threshold probabilities that are better suited for each subgroup.[40]

Going forward, when developing new machine learning classifiers and risk scores, we recommend consideration of equity[11] during design phase and model development phase. Although our post-hoc methods mitigated the FNR bias, other techniques like transfer learning during model training have been shown to improve performance for data-disadvantaged groups[41] and show promise. The Black subgroup in our chart review had greater severity of illness with more comorbidities as well as differences in age and risk for death. Features in the clinical notes from the Black subgroup may be weighted differently using transfer learning approaches that focus training across subgroups to produce a fair classifier. Future work should prioritize rigorous distributive justice techniques across all stages of machine learning research including study design, data collection—including manual screening rates across subgroups—model training, model development, and model deployment.[42] Our post-hoc results demonstrating differing patient characteristics across subgroups may have been identified sooner with collaboration between ethics experts and data scientists to address the disparities and define barriers prior to model training.

Our assessment of bias and fairness is not without limitations. First, this bias assessment addresses bias in the data while multiple sources of bias may persist. For example, selection bias and missingness due to underscreening[25]—which also may be attributable to implicit bias—in the external validation dataset may have impacted model performance. Similarly, the secular trends that Ciccarone and colleagues[12] identified may have introduced bias into the validation dataset due to the dominance of fentanyl opioids between 2017 and 2019. Further, with a blackbox CNN, we cannot fully interpret the associations underlying the model's predictions, but only estimate its features and weights with local surrogate models (ie, LIME). With the highly imbalanced data and the small number of positive cases, an intersectional analysis of bias and disparate impact was not possible.[43–45] Qualitative examinations of disparate impact and bias in the data may be informative. For example, as the rate of Hispanic/Latinx opioid-related overdose deaths continue to rise in Chicago,[26] ethnographic inquiries into opioid misuse and access to care within Hispanic/Latinx ethnic and racial subgroups and across broader racial subgroups could help identify opioid misuse treatment inequities.

CONCLUSION

Prior publications validating our opioid misuse classifier were focused on the predictive analytics with little insight into potential biases. We show in this study that these flawed sources of model explanation and interpretation perpetuate racism and the associated health inequities. Although similar features were present between subgroups for predicting opioid misuse, inequities were also present and may only be partially addressed with post-hoc recalibration. We showed differing risks for health outcomes between Black and White subgroups; so, it is not surprising that a one-size-fits all classifier underperforms in racial subgroups. Future work should focus on improvements to model debiasing techniques that are insensitive to fine-tuning and recalibration.

FUNDING

This research was supported in part by the following: (1) Agency for Healthcare Research & Quality Grant Number K12-HS-026385 (HT) and (2) National Institute for Drug Abuse R25DA035692 (HT); (3) National Institute of Drug Abuse Grant Number R01DA051464 (MA) and (4) National Institute on Alcohol Abuse and Alcoholism K23AA024503 (MA); (5) National Institute on Drug Abuse Grant Numbers R01DA04171 (NK), (6) UG1- DA-049467 (MA, NK), and (7) National Center for Advancing Translational Sciences Grant Number UL1-TR-002398 (MA, NK); (8) Dr. Matthew M. Churpeck declares patent pending (ARCD.P0535US.P2) for a risk model and research support from National Institute of General Medical Sciences R01GM123193; (9) National Library of Medicine R01LM012973 and R01LM010090 (DD).

AUTHOR CONTRIBUTIONS

HT conducted research, analysis, interpretation, and was the lead writer of the manuscript. In addition to contributing to editing, revising, and approving drafts, the coauthors contributed the following: BS conducted analysis, interpretation, and generated the figures for the manuscript. RB, CM, and SB conducted the subgroup FNR validation. DD conducted research and analysis on the development dataset. MC contributed to study design and interpretation. NK provided access to Rush data, interpretation, and content expertise in medical ethics and opioid misuse. MA designed the study, provided access to Loyola data, and conducted analysis, writing, and editing.

CONFLICT OF INTEREST STATEMENT

None declared.

DATA AVAILABILITY STATEMENT

The data underlying this article cannot be shared publicly. Our dataset is derived from patient electronic health records across 2 independent health systems and include protected health information. We cannot make the dataset publicly available due to regulatory and legal restrictions imposed by Rush University Medical Center and Loyola University Medical Center. Patient medical data is highly sensitive and, with quasi-identifiers such as race, ethnicity, and age, medical record data are reidentifiable when linked to other publicly available datasets. Should researchers who meet the criteria for access to this confidential data want to use our deidentified dataset to replicate the bias assessment and mitigation techniques, our Chief of Research Informatics at Rush (Dr. Casey Frankenberger) and Vice-Provost for Research at Loyola (Dr. Meharvan Singh) will serve as the points of contact outside our research team for inquiries about establishing a data use agreement and access to the patient data (cfranken@rush.edu, msingh@luc.edu) .

ACKNOWLEDGMENTS

The authors would like to acknowledge the Rush Department of Psychiatry Addiction Medicine Data Science Lab, the Substance Use Intervention Team, and Ketzel Feasley for their contributions to the publication of this manuscript.

REFERENCES

1. Burnside M, Crocket H, Mayo M, *et al.* Do-it-yourself automated insulin delivery: a leading example of the democratization of medicine. *J Diabetes Sci Technol* 2020; 14 (5): 878–82.
2. Allen B, Agarwal S, Kalpathy-Cramer J, *et al.* Democratizing AI. *J Am Coll Radiol* 2019; 16 (7): 961–3.
3. Gupta V, Roth H, Buch V, *et al.* Democratizing artificial intelligence in healthcare: a study of model development across 2 institutions incorporating transfer learning. arXiv[eess.IV], http://arxiv.org/abs/2009.12437, 2020, preprint: not peer reviewed.
4. Sharma B, Dligach D, Swope K, *et al.* Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients. *BMC Med Inform Decis Mak* 2020; 20 (1): 79.
5. Afshar M, Phillips A, Karnik N, *et al.* Natural language processing and machine learning to identify alcohol misuse from the electronic health record in trauma patients: development and internal validation. *J Am Med Inform Assoc* 2019; 26 (3): 254–61.
6. Afshar M, Joyce C, Dligach D, *et al.* Subtypes in patients with opioid misuse: a prognostic enrichment strategy using electronic health record data in hospitalized patients. *PLoS One* 2019; 14 (7): e0219717.
7. Bender EM, Friedman B. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Trans Asscoc Comput Linguist* 2018; 6: 587–604.
8. Benthall S, Haynes BD. Racial categories in machine learning. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. New York, NY: Association for Computing Machinery; 2019: 289–98.
9. Obermeyer Z, Powers B, Vogeli C, *et al.* Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447–53.
10. Dixon L, Li J, Sorensen J, *et al.* Measuring and mitigating unintended bias in text classification. In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. New York, NY: Association for Computing Machinery; 2018: 67–73.
11. Karnik NS, Afshar M, Churpek MM, *et al.* Structural disparities in data science: a prolegomenon for the future of machine learning. *Am J Bioeth* 2020; 20 (11): 35–7.
12. Ciccarone D. The triple wave epidemic: supply and demand drivers of the US opioid overdose crisis. *Int J Drug Policy* 2019; 71: 183–8.
13. Unick GJ, Ciccarone D. US regional and demographic differences in prescription opioid and heroin-related overdose hospitalizations. *Int J Drug Policy* 2017; 46: 112–9. doi:10.1016/j.drugpo.2017.06.003.
14. Netherland J, Hansen H. White opioids: pharmaceutical race and the war on drugs that wasn't. *Biosocieties* 2017; 12 (2): 217–38.
15. Hansen H, Netherland J. Is the prescription opioid epidemic a white problem? *Am J Public Health* 2016; 106 (12): 2127–9.
16. Persmark A, Wemrell M, Evans CR, *et al.* Intersectional inequalities and the US opioid crisis: challenging dominant narratives and revealing heterogeneities. *Crit Public Health* 2020; 30 (4): 398–414.
17. Roberts SCM, Nuru-Jeter A. Universal alcohol/drug screening in prenatal care: a strategy for reducing racial disparities? Questioning the assumptions. *Matern Child Health J* 2011; 15 (8): 1127–34.
18. Goedel WC, Shapiro A, Cerdá M, *et al.* Association of racial/ethnic segregation with treatment capacity for opioid use disorder in counties in the United States. *JAMA Netw Open* 2020; 3 (4): e203711.
19. Bauer GR, Lizotte DJ. Artificial intelligence, intersectionality, and the future of public health. *Am J Public Health* 2021; 111 (1): 98–100.
20. Saleiro P, Kuester B, Hinkson L, *et al.* Aequitas: A bias and fairness audit toolkit. arXiv [cs.LG], doi: http://arxiv.org/abs/1811.05577, 2018, preprint: not peer reviewed.
21. Ribeiro MT, Singh S, Guestrin C. "Why should I trust you?" Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 2016: 1135–44. doi: 10.1145/2939672.2939778.
22. Thompson HM, Hill K, Jadhav R, *et al.* The substance use intervention team: a preliminary analysis of a population-level strategy to address the opioid crisis at an Academic Health Center. *J Addict Med* 2019; 13 (6): 460–3.
23. Thompson HM, Faig W, VanKim NA, *et al.* Differences in length of stay and discharge destination among patients with substance use disorders: the effect of Substance Use Intervention Team (SUIT) consultation service. *PLoS One* 2020; 15 (10): e0239761.
24. Yudko E, Lozhkina O, Fouts A. A comprehensive review of the psychometric properties of the Drug Abuse Screening Test. *J Subst Abuse Treat* 2007; 32 (2): 189–98. doi:10.1016/j.jsat.2006.08.002.
25. Afshar M, Sharma B, Bhalla S, *et al.* External validation of an opioid misuse machine learning classifier in hospitalized adult patients. *Addict Sci Clin Pract* 2021; 16 (1): 19.
26. Office of Epidemiology. 2018 Chicago Opioid Overdose Data Brief. 2019.https://www.chicago.gov/content/dam/city/depts/cdph/CDPH/Healthy%20Chicago/ChicagoOpioid2018.pdf. Accessed June 27, 2021.
27. O'Donnell J, Gladden RM, Mattson CL, *et al.* Vital signs: characteristics of drug overdose deaths involving opioids and stimulants - 24 states and the district of Columbia, January-June 2019. *MMWR Morb Mortal Wkly Rep* 2020; 69 (35): 1189–97.
28. Castañeda SF, Garcia ML, Lopez-Gurrola M, *et al.* Alcohol use, acculturation and socioeconomic status among Hispanic/Latino men and women: The Hispanic Community Health Study/Study of Latinos. *PLoS One* 2019; 14 (4): e0214906.
29. Klugman M, Hosgood HD 3rd, Hua S, *et al.* A longitudinal analysis of nondaily smokers: the Hispanic Community Health Study/Study of Latinos (HCHS/SOL). *Ann Epidemiol* 2020; 49: 61–7.
30. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?" Explaining the predictions of any classifier. arXiv:1602.04938 [cs.LG], doi:10.1145/2939672.2939778, preprint: not peer reviewed.
31. Molnar C. Interpretable Machine Learning: A Guide for Making Black Box models Explainable. 2019. https://christophm.github.io/interpretable-ml-book/.
32. Schisterman EF, Perkins NJ, Liu A, *et al.* Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples. *Epidemiology* 2005; 16 (1): 73–81.
33. Huang Y, Li W, Macheret F, *et al.* A tutorial on calibration measurements and calibration models for clinical prediction models. *J Am Med Inform Assoc* 2020; 27 (4): 621–33.
34. van Walraven C, Austin PC, Jennings A, *et al.* A modification of the Elixhauser comorbidity measures into a point system for hospital death using administrative data. *Med Care* 2009; 47 (6): 626–33.
35. Institute of Medicine (US) Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care; Smedley BD, Stith AY, *et al.*, editors. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. In: *Racial and Ethnic Disparities in Diagnosis and Treatment: A Review of the Evidence and a Consideration of Causes*. Washington, DC: National Academies Press; 2003.
36. 2018 National Healthcare Quality and Disparities Report. Rockville, MD: Agency for Healthcare Research and Quality; 2020. https://www.ahrq.gov/research/findings/nhqrdr/nhqdr18/index.html.

37. Hoffman KM, Trawalter S, Axt JR, *et al.* Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proc Natl Acad Sci USA* 2016; 113 (16): 4296–301.
38. Ding Y, Tang S, Liao SG, *et al.* Bias correction for selecting the minimal-error classifier from many machine learning models. *Bioinformatics* 2014; 30 (22): 3152–8.
39. Park Y, Hu J, Singh M, *et al.* Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Netw Open* 2021; 4 (4): e213909.
40. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. *BMJ* 2016; 352: i6.
41. Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. *Nat Commun* 2020; 11 (1): 5131.
42. Rajkomar A, Hardt M, Howell MD, *et al.* Ensuring fairness in machine learning to advance health equity. *Ann Intern Med* 2018; 169 (12): 866–72.
43. Crenshaw K. Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics. *Univ Chic Leg Forum* 1989; 1989: 8.
44. Crenshaw K. Mapping the margins: intersectionality, identity politics, and violence against women of color. *Stanford Law Rev* 1991; 43 (6): 1241–99.
45. Bowleg L. The problem with the phrase women and minorities: intersectionality-an important theoretical framework for public health. *Am J Public Health* 2012; 102 (7): 1267–73.