Bias Assessment and Correction in Machine Learning Algorithms: A Use-Case in a Natural Language Processing Algorithm to Identify Hospitalized Patients with Unhealthy Alcohol Use

Marissa Borgese, MS[1], Cara Joyce, PhD[2], Emily E. Anderson, MPH, PhD[2], Matthew M. Churpek, MD, MPH, PhD[3], Majid Afshar, MD, MSCR[2,3]

[1]Loyola University Chicago Stritch School of Medicine, Maywood, IL; [2]Loyola University Chicago, Chicago, IL; [3]University of Wisconsin, Madison, WI

Abstract

Unhealthy alcohol use represents a major economic burden and cause of morbidity and mortality in the United States. Implementation of interventions for unhealthy alcohol use depends on the availability and accuracy of screening tools. Our group previously applied methods in natural language processing and machine learning to build a classifier for unhealthy alcohol use. In this study, we sought to evaluate and address bias through the use-case of our classifier. We demonstrated the presence of biased unhealthy alcohol use risk underestimation among Hispanic compared to Non-Hispanic White trauma inpatients, 18- to 44-year-old compared to 45 years and older medical/surgical inpatients, and Non-Hispanic Black compared to Non-Hispanic White medical/surgical inpatients. We further showed that intercept, slope, and concurrent intercept and slope recalibration resulted in minimal or no improvements in bias-indicating metrics within these subgroups. Our results exemplify the importance of integrating bias assessment early into the classifier development pipeline.

Introduction

An estimated 14 million United States adults met criteria for an alcohol use disorder in 2018, with nearly 1 in 10 adult alcohol users affected[1]. Unhealthy alcohol use represents a major economic burden and causal factor in cases of traumatic injury, liver disease, and other noncommunicable diseases resulting in death[2, 3]. Further, alcohol-related disorders consistently rank in the top-ten most common non-maternal diagnoses responsible for inpatient stays among adults under 44 years old[4] and individuals with alcohol-related disorders are more likely to return to the hospital within two weeks of being discharged compared to unaffected individuals[5]. Screening tools for unhealthy alcohol use can effectively identify individuals who will benefit from interventions that decrease alcohol use-related morbidity and mortality[6, 7]. Accordingly, the successful implementation of treatment depends on the availability and accuracy of such screening tools.

Unhealthy alcohol use screening in the hospital setting currently involves the use of a single screening question or a standardized questionnaire such as the Alcohol Use Disorders Identification Test (AUDIT)[8]. Adding self-report questionnaires into the clinical workflow requires additional time and resources that may limit their utility. Screening methods that utilize notes from the electronic health record (EHR) captured during routine care are promising alternative approaches to identify likely cases of unhealthy alcohol use[9].

For clinical decision support tools, there is growing concern about the potential for bias in machine learning (ML)-based automated approaches. There have been several notable publications discussing unintended bias in ML tools across several fields, with consequences ranging from underprediction of health risk in Black patients by a widely used commercial insurance algorithm[10] to increased error rate in speech recognition for Black speakers[11] and undervaluing of female job candidates[12]. In medicine, minimizing bias in ML-based clinical decision support is critical to avoiding downstream harm and the exacerbation of existing healthcare disparities.

Our group previously applied methods in natural language processing (NLP) and ML to build a classifier for unhealthy alcohol use from notes collected in the EHR that offers adequate sensitivity for screening patients in the acute care setting[9]. The classifier was trained on EHR notes from patients with a primary admission for trauma and has since been validated using data from non-trauma inpatient hospitalizations[13]. In this study, we seek to evaluate and address bias through the use-case of our previously developed unhealthy alcohol use NLP classifier. Our aims are the following: (1) to assess for bias in the NLP classifier via examination of subgroups of age, sex, and race/ethnicity; and (2) to determine if recalibration among subgroups affected by model bias can mitigate the bias in the NLP classifier's screening performance.

Methods

Unhealthy Alcohol Use Classifier

The unhealthy alcohol use NLP classifier previously published by our group[9] was developed using concept unique identifiers (CUIs) derived from linguistic processing of clinical notes, with AUDIT scores ≥5 and ≥8 points as the reference standard for unhealthy alcohol use in women and men, respectively. Clinical notes available from within 24 hours of presentation to the Emergency Department were scanned for Unified Medical Language System entity mentions, which were mapped to CUIs. Hyperparameters were tuned to maximize the area under the receiver operating characteristic curve (AUC ROC) using 10-fold cross-validation. The final classifier retained 16 CUI features and achieved an average AUC ROC of 0.78 (95% CI 0.68-0.89).

Patient Setting and Data

Our data consisted of two cohorts: one from the original development paper using trauma patient encounters in the Emergency Department (n=1,326) and one from the validation cohort of an independent group of hospitalized patients in the medical/surgical wards (n=999).

Our trauma dataset included consecutive patients seen at Loyola University Medical Center's (LUMC) Trauma Center who received screening for unhealthy alcohol use using the AUDIT questionnaire between April 2013 and November 2016. Results from the AUDIT were used as the reference standard for labeling cases of unhealthy alcohol use, with scores ≥5 and ≥8 points indicating unhealthy alcohol use in women and men, respectively[8]. The trauma cohort reflects the pooled training and internal validation data used for the initial development and validation of the unhealthy alcohol use NLP classifier[9].

Our inpatient dataset consisted of a convenience sample of patients who presented to LUMC for non-trauma inpatient hospitalizations between January 2007 and September 2017. Unhealthy alcohol use was identified via chart review by a trained annotator following standardized criteria previously described[13] with an oversampling of at-risk patients to provide a more balanced dataset for evaluating cases. The inpatient cohort reflect the data used for external validation of the NLP classifier[13].

Predicted probabilities of unhealthy alcohol use were generated for all encounters in both cohorts, and the optimal cutoffs were determined using the Youden index maximization method. The optimal predicted probability cutoffs for identifying unhealthy alcohol use were ≥0.46 and ≥0.53 for the trauma and inpatient cohorts, respectively. In the trauma cohort, the classifier had a sensitivity and specificity of 61% (95% CI 56%-67%) and 78% (95% CI 75%-80%), respectively. In the external validation inpatient cohort, the classifier had a sensitivity and specificity of 86% (95% CI 83%-89%) and 82% (95% CI 78%-85%), respectively. Descriptive statistics for patient characteristics, including age, sex, ethnicity, and race, were presented for both cohorts. All analyses were performed using R version 3.6.0 (R Core Team, 2019).

Bias Assessment

We assessed for bias in the trauma and inpatient cohorts independently and by age group, sex, and race/ethnicity. For the purposes of bias assessment and correction, age was divided into two groups *a priori* based on census age groups and sample sizes: 18 to 44 years old and 45 years and older. Race/ethnicity was divided into three groups: Hispanic, Non-Hispanic Black, and Non-Hispanic White. The reference standard labels and predicted labels for unhealthy alcohol use (present/absent) were used to calculate the number of false positives (FP), true positives (TP), false negatives (FN) and true negatives (TN) within each cohort and within subgroups of each cohort. We then calculated the bias assessment metrics of interest, which included false discovery rate (FDR; FP/[FP+TP]), false positive rate (FPR; FP/[FP+TN]), false omission rate (FOR; FN/[FN+TN]), and false negative rate (FNR; FN/[FN+TP]). All bias metrics were calculated as described by Saleiro, Kuester[14] and 95% exact binomial confidence intervals were calculated for all metrics.

High FDR and/or FPR values were used as indicators of bias towards overestimation of risk within subgroups; high FOR and/or FNR values were used as indicators of bias towards underestimation of risk within subgroups. Combinations of high FDR and/or FPR values with high FOR and/or FNR values were interpreted as overall reduced model accuracy within subgroups.

Bias Correction

For the purposes of developing an adequate screening tool with the NLP classifier, we focused our bias correction efforts on maintaining high sensitivity with few false negative results, so subgroups with high FOR and/or FNR values were targeted. Model calibration of intercept and slope was assessed using calibration plots in accordance with the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines[15]. Model recalibration was implemented for each cohort subgroup that had evidence of poor calibration. Predicted probabilities within affected subgroups were used to calculate linear predictors in a logistic regression classifier, which were subsequently used for model recalibration via intercept re-estimation, slope re-estimation, and concurrent intercept and slope re-estimation. These recalibration methods were chosen to minimize the number of re-estimated parameters, control familywise error rate, and to avoid overfitting given the relatively small sample sizes available among select subgroups within both cohorts[16]. The original NLP classifier and all three recalibrated classifiers were compared via calibration plots and scaled Brier scores with bootstrapped 95% confidence intervals (1,000 iterations). A model's scaled Brier score reflects the mean square error of probability predictions, or Brier score, scaled by its maximum score such that values range from 0% to 100% and 100% indicates optimal performance[17]. The bias-indicating metrics were also recalculated for affected subgroups using predicted probabilities and a predicted probability threshold determined via Youden index maximization from each of the three recalibration methods.

Results

Patient Characteristics

Our trauma cohort consisted of 1,326 encounters from 1,309 patients and our inpatient cohort consisted of 999 encounters from 856 patients. Age at first encounter, sex, ethnicity, and race distributions were similar in the trauma and inpatient cohorts (**Table 1**).

Table 1. Patient characteristics by cohort.

Parameters	Trauma cohort (n=1,309)	Inpatient Cohort (n=856)
Age, years, median (IQR)	45 (28-62)	50 (39-59)
Sex, male, n (%)	921 (70)	567 (66)
Hispanic, n (%)	231 (18)	109 (13)
Race, n (%)		
American Indian	1 (0)	2 (0)
Asian	10 (1)	1 (0)
Black	323 (25)	261 (30)
White	778 (59)	502 (59)
Other	188 (14)	86 (10)

IQR = interquartile range.

Trauma Cohort Bias Assessment

The case-rate of unhealthy alcohol use was 23% (n=305) across all encounters in the trauma cohort (**Table 2**). The FPR and FOR were higher among patients 18 to 44 years old compared to patients 45 years and older, indicating reduced model accuracy among adults 44 years and younger in the trauma cohort. The FPR and FOR were also higher among male patients compared to female patients, indicating reduced model accuracy among male patients in the trauma cohort. The FOR was higher among Hispanic patients (FOR=0.22, 95% CI 0.16-0.30) compared to Non-Hispanic White patients (FOR=0.11, 95% CI 0.08-0.14), indicating biased underestimation of unhealthy alcohol use risk among Hispanic patients compared to Non-Hispanic White patients in the trauma cohort by the NLP classifier.

Inpatient Cohort Bias Assessment

The case-rate of unhealthy alcohol use was 58% (n=579) in the inpatient cohort (**Table 3**). The FOR and FNR were higher among patients 18 to 44 years old (FOR=0.32, 95% CI 0.25-0.40; FNR=0.20, 95% CI 0.15-0.26) compared to patients 45 years and older (FOR=0.12, 95% CI 0.09-0.17; FNR=0.10, 95% CI 0.07-0.14), indicating biased underestimation of unhealthy alcohol use risk among adults 44 years and younger compared to adults 45 years and older in the inpatient cohort. The FOR and FNR were also higher among Non-Hispanic Black patients (FOR=0.28,

Table 2. Bias report for the trauma cohort.

Population	n	Unhealthy Alcohol Use Prevalence	FDR	FPR	FOR	FNR
All encounters	1,326	0.23	0.55 (0.50-0.60)	0.22 (0.20-0.25)	0.13 (0.11-0.15)	0.39 (0.33-0.44)
Age group						
18 to 44 years	654	0.28	0.60 (0.53-0.66)	**0.33** **(0.29-0.37)**	**0.20** **(0.16-0.24)**	0.43 (0.36-0.50)
45 years and older	672	0.18	0.47 (0.39-0.55)	0.13 (0.11-0.16)	0.08 (0.06-0.10)	0.33 (0.24-0.42)
Sex						
Female	394	0.16	0.56 (0.45-0.66)	0.15 (0.11-0.19)	0.08 (0.05-0.11)	0.37 (0.25-0.50)
Male	932	0.26	0.55 (0.49-0.60)	**0.26** **(0.23-0.30)**	**0.16** **(0.13-0.19)**	0.39 (0.33-0.46)
Race/ethnicity						
Hispanic	232	0.33	0.46 (0.35-0.58)	0.22 (0.16-0.30)	**0.22** **(0.16-0.30)**	0.46 (0.35-0.58)
Non-Hispanic Black	324	0.21	0.65 (0.55-0.73)	0.29 (0.23-0.35)	0.14 (0.09-0.19)	0.42 (0.30-0.55)
Non-Hispanic White	698	0.21	0.54 (0.46-0.61)	0.19 (0.16-0.23)	0.11 (0.08-0.14)	0.36 (0.29-0.45)

FDR=false discovery rate; FPR=false positive rate; FOR=false omission rate; FNR=false negative rate.

95% CI 0.21-0.37; FNR=0.21, 95% CI 0.15-0.27) compared to Non-Hispanic White patients (FOR=0.13, 95% 0.09-0.18; FNR=0.10, 95% CI 0.07-0.14), indicating underestimation of risk among Non-Hispanic Black patients compared to Non-Hispanic White patients in the inpatient cohort.

Table 3. Bias report for the inpatient cohort.

Population	n	Unhealthy Alcohol Use Prevalence	FDR	FPR	FOR	FNR
All encounters	999	0.58	0.13 (0.11-0.16)	0.18 (0.15-0.22)	0.19 (0.16-0.23)	0.14 (0.11-0.17)
Age group						
18 to 44 years	360	0.66	0.10 (0.06-0.15)	0.17 (0.11-0.25)	**0.32** **(0.25-0.40)**	**0.20** **(0.15-0.26)**
45 years and older	639	0.54	0.15 (0.12-0.20)	0.19 (0.15-0.24)	0.12 (0.09-0.17)	0.10 (0.07-0.14)
Sex						
Female	343	0.53	0.14 (0.09-0.20)	0.16 (0.10-0.22)	0.18 (0.12-0.24)	0.16 (0.11-0.22)
Male	656	0.61	0.13 (0.10-0.17)	0.20 (0.15-0.26)	0.20 (0.16-0.26)	0.13 (0.10-0.17)
Race/ethnicity						
Hispanic	117	0.59	0.18 (0.09-0.29)	0.25 (0.14-0.40)	0.27 (0.15-0.41)	0.19 (0.10-0.30)
Non-Hispanic Black	306	0.60	0.15 (0.10-0.21)	0.21 (0.14-0.29)	**0.28** **(0.21-0.37)**	**0.21** **(0.15-0.27)**
Non-Hispanic White	538	0.56	0.13 (0.09-0.17)	0.16 (0.12-0.22)	0.13 (0.09-0.18)	0.10 (0.07-0.14)

FDR=false discovery rate; FPR=false positive rate; FOR=false omission rate; FNR=false negative rate.

Trauma Cohort Bias Correction

For the FOR and FNR metrics, the strongest bias in the trauma cohort with significant disparity was between the Hispanic and Non-Hispanic White subgroups. The original calibration primarily demonstrated overestimation of risk across the quintiles of predicted probabilities (**Figure 1**). Efforts at recalibration show improvement with movement towards the loess curve for perfect calibration. The concurrent intercept and slope re-estimation model resulted in the most significant improvement (**Figure 1**). The scaled Brier scores also reflected improvements in accuracy with intercept re-estimation (12%, 95% CI 9%-14%), slope re-estimation (15%, 95% CI 10%-21%), and concurrent intercept and slope re-estimation (17%, 95% CI 12%-21%) over the original model (7%, 95% CI 2%-10%).

The optimal predicted probability cutoffs were ≥0.33, ≥0.37, and ≥0.33 for the intercept re-estimation, slope re-estimation, and concurrent intercept and slope re-estimation models in the Hispanic subgroup, respectively. Despite improvements in model accuracy with recalibration, biased underestimation of unhealthy alcohol use risk (as indicated by a high FOR) persisted in the intercept re-estimation (FOR=0.21, 95% CI 0.15-0.29), slope re-estimation (FOR=0.21, 95% CI 0.15-0.28), and concurrent intercept and slope re-estimation (FOR=0.21, 95% CI 0.15-0.29) models compared to the original model (FOR=0.22, 95% CI 0.16-0.30). Minimal changes in FDR, FPR, and FNR were observed after recalibration.

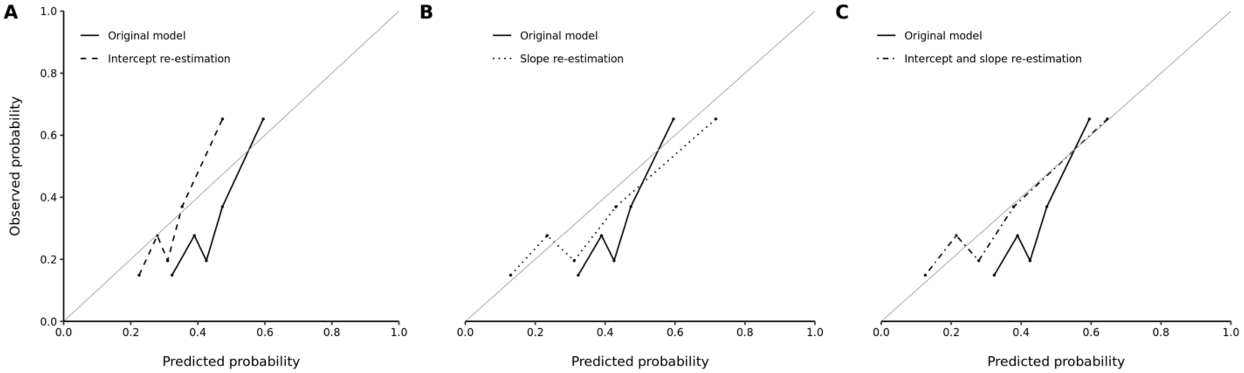

Figure 1. Observed versus predicted probability of unhealthy alcohol use among Hispanic patients in the trauma cohort according to the (A) original versus intercept re-estimation model, (B) original versus slope re-estimation model, and (C) original versus concurrent intercept and slope re-estimation model. Five bins were used for each graph.

Inpatient Cohort Bias Correction

For the FOR and FNR metrics, the strongest bias in the inpatient cohort with significant disparity was between the age groups. Our recalibration methods failed to produce noticeable improvements in predictive accuracy of the NLP classifier within the 18- to 44-year-old subgroup of the inpatient cohort (**Figure 2**). The scaled Brier scores also indicated a lack of improvement in model accuracy with intercept re-estimation (41%, 95% CI 37%-44%), slope re-estimation (44%, 95% CI 40%-49%), and concurrent intercept and slope re-estimation (44%, 95% CI 40%-49%) over the original model (41%, 95% CI 37%-44%).

The optimal predicted probability cutoffs were ≥0.47, ≥0.43, and ≥0.43 for the intercept re-estimation, slope re-estimation, and concurrent intercept and slope re-estimation models in the 18- to 44-year-old subgroup, respectively. Using the recalibrated predicted probabilities and cutoffs, there was an equal, moderate decrease in FNR (FNR=0.12, 95% CI 0.08-0.17) across all recalibration methods when compared to the original model (FNR=0.20, 95% CI 0.15-0.26). We observed minimal changes in FDR, FPR, and FOR with recalibration.

As was the case in the 18- to 44-year-old subgroup, no noteworthy improvements in model accuracy were observed with model recalibration within the Non-Hispanic Black subgroup of the inpatient cohort (**Figure 3**). The scaled Brier scores similarly indicated a lack of improvement in model accuracy with intercept re-estimation (41%, 95% CI 37%-44%), slope re-estimation (43%, 95% CI 38%-48%), and concurrent intercept and slope re-estimation (43%, 95% CI 39%-48%) over the original model (40%, 95% CI 37%-44%).

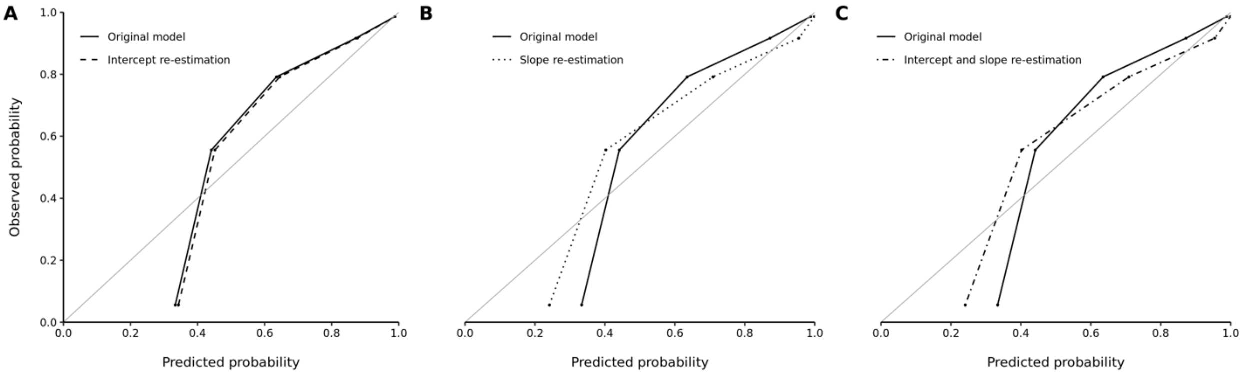

Figure 2. Observed versus predicted probability of unhealthy alcohol use among 18- to 44-year-old patients in the inpatient cohort according to the (A) original versus intercept re-estimation model, (B) original versus slope re-estimation model, and (C) original versus concurrent intercept and slope re-estimation model. Five bins were used for each graph.

The optimal predicted probability cutoffs were ≥0.55, ≥0.63, and ≥0.59 for the intercept re-estimation, slope re-estimation, and concurrent intercept and slope re-estimation models in the Non-Hispanic Black subgroup, respectively. We further found that biased underestimation of unhealthy alcohol use risk persisted across the recalibrated models with regards to both FOR and FNR compared to the original model. There were also minimal changes in FDR and FPR with recalibration.

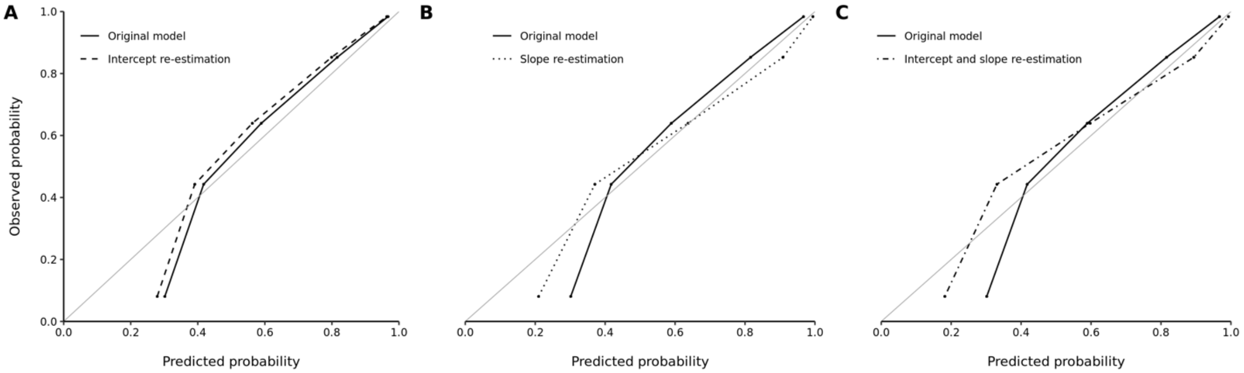

Figure 3. Observed versus predicted probability of unhealthy alcohol use among Non-Hispanic Black patients in the inpatient cohort according to the (A) original versus intercept re-estimation model, (B) original versus slope re-estimation model, and (C) original versus concurrent intercept and slope re-estimation model. Five bins were used for each graph.

Discussion

Our study demonstrates biased underprediction of unhealthy alcohol use by our NLP classifier for Hispanic patients compared to Non-Hispanic White patients admitted for trauma. For non-trauma hospitalizations, we also demonstrate biased underprediction of unhealthy alcohol use for adult patients under 45 years old compared to adults 45 years and older and Non-Hispanic Black patients compared to Non-Hispanic White patients. Moreover, we show that the greatest improvement in classifier accuracy with recalibration is achieved via concurrent intercept and slope recalibration within the Hispanic subgroup of the trauma cohort. Finally, none of the recalibration methods we implemented adequately addressed the risk underprediction disparities seen across age groups and racial/ethnic identities.

Both of our study cohorts have a median age between 40 and 50 years old and are predominantly male and Non-Hispanic White. The demographic composition of the training cohort is likely responsible for some of the bias seen in our NLP classifier. Our trauma cohort represents all patients who presented to the LUMC Emergency Department for trauma over a span of three years and thus is a better representation of true unhealthy alcohol use prevalence than our inpatient cohort, which includes an oversampling of encounters with a high likelihood of associated unhealthy alcohol use. Whole-cohort bias metrics are similar across the two cohorts except for the high FDR and FNR in the trauma cohort; these values are likely driven by the lower sensitivity and thus smaller number of TPs in the trauma cohort.

We were able to identify disparities in over- and underprediction of unhealthy alcohol use risk by examining FDR, FPR, FOR, and FNR resulting from our NLP classifier across demographic subgroups. Such algorithmic biases can be introduced in several places along the model development pipeline, including through the training data, model design, and threshold selection. The data used in the development of our NLP classifier resulted from a consecutive sample of patients seen over three years, though approximately half of the eligible patients in this time were not screened for unhealthy alcohol use via the AUDIT[9]. Deviations from true consecutive sampling can result in selection bias due to underrepresentation of certain patient populations; this is particularly problematic when underrepresentation of minority groups is potentially involved. Measurement bias is also a concern when developing NLP tools, especially given existing evidence that implicit racial biases can influence physician language[18] and the knowledge that biased physician language can be perpetuated through NLP features. Biases can further be perpetuated by using a single risk threshold, which can result in a lack of assistive measures for patient populations with systematic risk underprediction or an excess of punitive measures for patient populations with systematic risk overprediction.

Through our attempts to mitigate the biases identified in our unhealthy alcohol use classifier, we found that model recalibration was insufficient to address disparities in risk underprediction-indicating metrics across demographic subgroups. Alternative methods for minimizing bias in classification models can be used, starting with measures to reduce selection and measurement biases during data collection. Additional methods focus specifically on minimizing bias among minority populations through improvements in accuracy, including oversampling with subsequent weighting of minority population data and transfer learning[19]. Our approach of *post hoc* recalibration within subgroups with biased risk underprediction resulted in some modest improvements in classifier accuracy. However, this method failed to produce improvements in FOR and FNR that would translate to positive effects in clinical practice. Our classifier in its current form stands to recommend assistive unhealthy alcohol use interventions to Hispanic trauma patients less frequently than to Non-Hispanic White trauma patients. Our results likewise demonstrate the value of bias assessment across patient subgroups rather than solely relying on global accuracy metrics of classifier models.

Our study has several limitations. We were unable to compare classifier bias between cohorts due to the use of two different gold standards for identifying unhealthy alcohol use. We were also unable to assess whether using different, sex-dependent AUDIT cutoffs for reference standard unhealthy alcohol use labeling in the trauma cohort introduced or masked classifier bias. Small sample sizes within some subgroups required us to collapse groups down, as in the age groups, or forego analysis altogether, as in the American Indian and Asian racial groups. Counts of FPs, TPs, FNs, and TNs were also limited (<25) in four instances. Further, we were unable to assess for bias by gender identity as this data was not available.

We demonstrated the presence of biased unhealthy alcohol use risk underestimation by our NLP classifier among Hispanic compared to Non-Hispanic White trauma inpatients, 18- to 44-year-old compared to 45 years and older medical/surgical inpatients, and Non-Hispanic Black compared to Non-Hispanic White medical/surgical inpatients. We further showed that intercept, slope, and concurrent intercept and slope recalibration resulted in minimal or no improvements in bias-indicating metrics within these subgroups. In summary, we detected bias in our NLP classifier and were unable to adequately address this bias through *post ho*c recalibration methods. Our results exemplify the importance of integrating bias assessment early into the classifier algorithm development pipeline, ideally in collaboration with health equity researchers and with consideration of the complex nature of bias as it relates to structural health disparities[20].

References

1. Center for Behavioral Health Statistics and Quality SAaMHSA. Key substance use and mental health indicators in the united states: Results from the 2018 national survey on drug use and health. 2019.
2. World Health Organization. Global status report on alcohol and health 2018. 2018.
3. Moon AM, Yang JY, Barritt ASt, Bataller R, Peery AF. Rising mortality from alcohol-associated liver disease in the united states in the 21st century. Am J Gastroenterol. 2019.
4. Healthcare Cost and Utilization Project (HCUP). Most common diagnoses for inpatient stays. In: Agency for Healthcare Research and Quality, editor. 2019.

5. Montoy JCC, Tamayo-Sarver J, Miller GA, Baer AE, Peabody CR. Predicting emergency department "bouncebacks": A retrospective cohort analysis. West J Emerg Med. 2019;20(6):865-74.
6. Rogal S, Youk A, Zhang H, et al. Impact of alcohol use disorder treatment on clinical outcomes among patients with cirrhosis. Hepatology. 2019.
7. Charlet K, Heinz A. Harm reduction-a systematic review on effects of alcohol reduction on physical and mental symptoms. Addict Biol. 2017;22(5):1119-59.
8. U.S. Department of Health and Human Services, National Institutes of Health, National Institue on Alcohol Abuse and Alcoholism. Helping patients who drink too much: A clinician's guide. 2005.
9. 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.
10. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447-53.
11. Koenecke A, Nam A, Lake E, et al. Racial disparities in automated speech recognition. Proc Natl Acad Sci U S A. 2020;117(14):7684-9.
12. Dastin J. Amazon scraps secret ai recruiting tool that showed bias against women. Business News [Internet]. 2018. Available from: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G.
13. To D, Sharma B, Karnik N, Joyce C, Dligach D, Afshar M. Validation of an alcohol misuse classifier in hospitalized patients. Alcohol. 2019.
14. Saleiro P, Kuester B, Stevens A, et al. Aequitas: A bias and fairness audit toolkit. arXiv preprint arXiv:181105577. 2018.
15. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod): The tripod statement. Ann Intern Med. 2015;162(1):55-63.
16. Vergouwe Y, Nieboer D, Oostenbrink R, et al. A closed testing procedure to select an appropriate method for updating prediction models. Stat Med. 2017;36(28):4529-39.
17. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
18. Hagiwara N, Slatcher RB, Eggly S, Penner LA. Physician racial bias and word use during racially discordant medical interactions. Health Commun. 2017;32(4):401-8.
19. Gao Y, Cui Y. Deep transfer learning for reducing health care disparities arising from biomedical data inequality. Nat Commun. 2020;11(1):5131.
20. Karnik NS, Afshar M, Churpek MM, Nunez-Smith M. Structural disparities in data science: A prolegomenon for the future of machine learning. Am J Bioeth. 2020;20(11):35-7.