OPEN

Racial Differences in Accuracy of Predictive Models for High-Flow Nasal Cannula Failure in COVID-19

Philip Yang, MD, MSc[1]

Ismail A. Gregory, MD[1]

Chad Robichaux, MPH[2]

Andre L. Holder, MD, MSc[1]

Greg S. Martin, MD, MSc[1]

Annette M. Esper, MD, MSc[1]

Rishikesan Kamaleswaran, PhD[2,3]

Judy W. Gichoya, MD[4]

Sivasubramanium V. Bhavani, MD, MS[1]

OBJECTIVES: To develop and validate machine learning (ML) models to predict high-flow nasal cannula (HFNC) failure in COVID-19, compare their performance to the respiratory rate-oxygenation (ROX) index, and evaluate model accuracy by self-reported race.

DESIGN: Retrospective cohort study.

SETTING: Four Emory University Hospitals in Atlanta, GA.

PATIENTS: Adult patients hospitalized with COVID-19 between March 2020 and April 2022 who received HFNC therapy within 24 hours of ICU admission were included.

INTERVENTIONS: None.

MEASUREMENTS AND MAIN RESULTS: Four types of supervised ML models were developed for predicting HFNC failure (defined as intubation or death within 7 d of HFNC initiation), using routine clinical variables from the first 24 hours of ICU admission. Models were trained on the first 60% ($n = 594$) of admissions and validated on the latter 40% ($n = 390$) of admissions to simulate prospective implementation. Among 984 patients included, 317 patients (32.2%) developed HFNC failure. eXtreme Gradient Boosting (XGB) model had the highest area under the receiver-operator characteristic curve (AUROC) for predicting HFNC failure (0.707), and was the only model with significantly better performance than the ROX index (AUROC 0.616). XGB model had significantly worse performance in Black patients compared with White patients (AUROC 0.663 vs. 0.808, $p = 0.02$). Racial differences in the XGB model were reduced and no longer statistically significant when restricted to patients with nonmissing arterial blood gas data, and when XGB model was developed to predict mortality (rather than the composite outcome of failure, which could be influenced by biased clinical decisions for intubation).

CONCLUSIONS: Our XGB model had better discrimination for predicting HFNC failure in COVID-19 than the ROX index, but had racial differences in accuracy of predictions. Further studies are needed to understand and mitigate potential sources of biases in clinical ML models and to improve their equitability.

KEYWORDS: acute respiratory failure; COVID-19; high-flow nasal cannula; machine learning

Heated and humidified high-flow nasal cannula (HFNC) has been a key component of noninvasive respiratory support for acute hypoxic respiratory failure due to COVID-19 (1–3). Nevertheless, between 40% and 76% of patients fail HFNC treatment and eventually require invasive mechanical ventilation (IMV) (4–7). There are uncertainties regarding risk factors for HFNC failure and the optimal threshold for initiating IMV (4, 5). As such, there is a need for accurate predictive models of HFNC treatment failure that

DOI: 10.1097/CCE.0000000000001059

KEY POINTS

Question: Are predictive models for high-flow nasal cannula treatment failure equitably accurate among patients of different races?

Findings: Supervised machine learning (ML) models, such as eXtreme Gradient Boosting (XGB) model, were developed with better discrimination for predicting HFNC failure in patients with COVID-19 than the respiratory rate-oxygenation index. However, the XGB model had worse accuracy in Black patients compared with White patients.

Meaning: Mitigation of potential biases in measurements, missing data and imputation strategies, and intubation thresholds may be needed to ensure equitable performance and application of clinical ML models.

may be helpful for risk stratification and allocation of healthcare resources (7, 8). However, the utility of clinical predictive models is partially limited by ongoing concern of differential performance in patients of different subgroups, which increases the risk of perpetuating healthcare disparities (9–15). Given the racial differences in rates of occult hypoxemia (16–18) and mechanical ventilation (19), there is an opportunity to understand racial differences in the performance of models that predict HFNC treatment failure.

The respiratory rate-oxygenation (ROX) index is a well-studied scoring system that has good discrimination in predicting the need for IMV after HFNC treatment in patients with pneumonia (20, 21) and COVID-19 (6–8, 22–26). The ROX index is the ratio of oxygen saturation to the Fio_2 and respiratory rate (ROX = [$\mathrm{Spo}_2/\mathrm{Fio}_2$]/RR) (20, 21). It is currently the only validated predictive model for HFNC treatment failure, and novel machine learning (ML) methods have not yet been applied to developing predictive models for HFNC failure. However, there is a need to investigate whether the ROX index performs equally well in Black and White patients, given that pulse oximetry, one of the components of the ROX index, has worse accuracy in Black patients (16–18). Furthermore, there is a need to evaluate whether novel ML models demonstrate racial differences in accuracy of predictions (9, 15, 27) due to potential biases in measurements (e.g., pulse oximetry) (16–18), missingness of data (28, 29), and/or intubation thresholds (19).

The aims of the study were: 1) to develop and validate an ML model to predict HFNC treatment failure, using available electronic medical record (EMR) data from hospitalized COVID-19 patients, 2) to compare the ML model performance to that of the ROX index, and 3) to evaluate potential racial differences in the accuracy of the ML models and the ROX index. We hypothesized that the ML model would predict HFNC failure in patients with COVID-19 with higher accuracy than the ROX index, and that both the ML model and the ROX index would have differences in accuracy by race.

MATERIALS AND METHODS

Patient Setting

This retrospective cohort study included adult patients 18 years old or older who were: 1) admitted to one of four Emory University Hospitals with a diagnosis of COVID-19 between March 1, 2020, and April 30, 2022, 2) started on HFNC therapy within 24 hours of ICU admission, and 3) remained on HFNC for at least 6 hours (full inclusion and exclusion criteria are detailed in **eTable 1**, http://links.lww.com/CCX/B319). Pertinent clinical data were extracted from the Emory Healthcare Clinical Data Warehouse (which imported clinical data from Cerner, North Kansas City, Missouri), deidentified, and made available for analyses. The study was approved by the institutional review board (IRB) at Emory University (STUDY00001627, "COVID-19 Subphenotypes," approved October 15, 2020) and was conducted in accordance with the ethical standards of Emory University IRB and the Helsinki Declaration of 1975. Based on general impracticability and minimal harm, waiver of consent was granted by the Emory University IRB.

Study Outcome

The primary outcome was HFNC failure, defined as requiring IMV or death up to 7 days since initiation of HFNC therapy. The 7-day window was chosen based on definitions used in prior studies and typical ranges of time to intubation and/or death reported in prior studies of HFNC failure (6, 22–24), as well as to reduce the likelihood of confounding factors contributing to delayed clinical deterioration.

Data Processing

We developed predictive models using EMR data available in the first 24 hours of admission, which included demographics, medical comorbidities, vital signs, laboratory values, medications, and bolus IV fluids administered; the full list of variables with missingness is reported in **eTable 2** (http://links.lww.com/CCX/B319). Missing values were imputed with the population median/mode for the primary analysis. In sensitivity analyses for evaluating potential biases in missingness and imputation, we also tested: 1) excluding patients with missing arterial blood gas (ABG) data, then applying median/mode imputation for other variables, 2) imputation with predictive mean matching (PMM) for all variables, and 3) leaving missing data as missing. For calculating the ROX index, missing values for $\mathrm{Sp{\scriptstyle O}}_2$ and RR were imputed using fill-forward imputation, and missing values for $\mathrm{F{\scriptstyle IO}}_2$ were imputed using fill-forward-then-backward imputation during the duration of HFNC treatment. The worst (lowest) ROX index during the duration of HFNC treatment in the first 24 hours was used to predict HFNC failure and to keep the input data consistent with the ML models.

Race was categorized as "Asian," "Black," "Other," or "White" according to the self-reported race information recorded in the EMR data. Self-identified ethnicity (i.e., Hispanic or non-Hispanic) was not considered in defining the race categories because of inconsistent documentation. Race and ethnicity were not used as predictor variables in any of the models, except only to perform stratified analyses by race.

Model Development and Validation

The first 60% of hospitalizations (March 1, 2020, to February 28, 2021; *n* = 594) were assigned to the training cohort, and latter 40% of hospitalizations (March 1, 2021, to April 30, 2022; *n* = 390) were assigned to the validation cohort. The dataset was split temporally to simulate prospective implementation of the models. Four supervised learning models including logistic regression (LR), eXtreme Gradient Boosting (XGB), k-nearest neighbors (KNNs), and support vector machine (SVM) models to predict HFNC failure were developed using the training cohort and tested on the validation cohort. Five-fold cross-validation was performed in the training set to tune model hyperparameters to maximize the area under the receiver-operator characteristic curve (AUROC) for all models. All models were developed following the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis and related guidelines (30, 31).

Model Performance Evaluation

Model performance of the ML models for predicting HFNC failure was compared with the ROX index using AUROC. For the models that performed significantly better than the ROX index, variable importance in model prediction was examined using SHapley Additive exPlanations (SHAP) plots. Model validations were also stratified by race, sex, and age group (< 65 vs. ≥ 65 yr old); the stratified analyses by race were restricted only to Black and White patients due to low numbers of patients in other race categories. Significant differences in model performance between subgroups were evaluated by generating calibration belts (32) and in sensitivity analyses.

Statistical Analysis

Patient characteristics and clinical variables were compared using t-test or Wilcoxon rank-sum test for continuous variables and chi-square or Fisher exact test for categorical variables, as appropriate. The models were compared by comparing the AUROC and 95% CIs generated from bootstrapping. All analyses were performed in R, v.4.2.0 (R Foundation for Statistical Computing, Vienna, Austria; used packages listed in **eTable 3**, http://links.lww.com/CCX/B319), and *p* value of less than 0.05 was used for statistical significance.

RESULTS

Patient Characteristics

There were 984 patients who satisfied the study inclusion/exclusion criteria and were included in the analysis. Of these, 317 patients (32.2%) experienced HFNC failure; their clinical characteristics are summarized in **Table 1** and **eTable 4** (http://links.lww.com/CCX/B319). Patients who failed HFNC therapy were older than those who did not fail (median [interquartile

TABLE 1.
Patient Characteristics and Clinical Outcomes, by High-Flow Nasal Cannula Treatment Failure Status

Characteristics	Overall (*n* = 984)	Failure (*n* = 317) (32.2%)	Nonfailure (*n* = 667) (67.8%)	*p*
Age (yr), median (IQR)	62 (51–73)	65 (55–76)	61 (50–71)	< 0.01
Sex, *n* (%)				
Female	444 (45.1%)	147 (46.4%)	297 (44.5%)	0.63
Male	540 (54.9%)	170 (53.6%)	370 (55.5%)	
Race, *n* (%)				
Asian	43 (4.4%)	14 (4.4%)	29 (4.4%)	0.89
Black	500 (50.8%)	163 (51.4%)	337 (50.5%)	
Other	111 (11.3%)	32 (10.1%)	79 (11.8%)	
White	330 (33.5%)	108 (34.1%)	222 (33.3%)	
Treatments received in the first 24 hr, *n* (%)				
Dexamethasone	780 (79.3%)	241 (76.0%)	539 (80.8%)	0.10
Remdesivir	505 (51.3%)	143 (45.1%)	362 (54.3%)	0.01
Norepinephrine	102 (10.4%)	64 (20.2%)	38 (5.7%)	< 0.01
Epinephrine	3 (0.3%)	1 (0.3%)	2 (0.3%)	1.00
Vasopressin	22 (2.2%)	10 (3.2%)	12 (1.8%)	0.25
Bolus IV fluids in the first 24 hr (mL), median (IQR)	750 (500–1250)	750 (500–1250)	750 (250–1250)	0.11
Required noninvasive ventilation, *n* (%)	204 (20.7%)	105 (33.1%)	99 (14.8%)	< 0.01
Required invasive mechanical ventilation, *n* (%)	404 (41.1%)	288 (90.9%)	116 (17.4%)	< 0.01
All-cause mortality, *n* (%)	190 (19.3%)	141 (44.5%)	49 (7.4%)	< 0.01
Time from admission to HFNC initiation (hr), median (IQR)	3.8 (0.8–11.5)	2.7 (0.7–8.9)	4.6 (0.9–12.3)	0.01
HFNC therapy duration (hr), median (IQR)	156 (72–280)	179 (62–379)	151 (79–249)	0.04
Hospital length of stay (hr), median (IQR)	256 (164–463)	409 (213–712)	221 (160–354)	< 0.01
Worst respiratory rate-oxygenation index in the first 24 hr, mean (SD)	6.35 ± 4.54	5.54 ± 4.36	7.11 ± 4.80	< 0.01
Sequential Organ Failure Assessment score at the time of HFNC initiation, mean (SD)	5.21 ± 2.89	7.21 ± 2.69	4.27 ± 2.47	< 0.01

HFNC = high-flow nasal cannula, IQR = interquartile range.

range (IQR)] 65 yr [55–76] vs. 61 yr [50–71], $p < 0.01$), but the two groups had otherwise comparable demographic characteristics. The failure group had a higher proportion of patients who required norepinephrine infusion in the first 24 hours (20.2% vs. 5.7%, $p < 0.01$), higher proportion requiring noninvasive ventilation (NIV; 33.1% vs. 14.8%, $p < 0.01$) during the hospitalization, and higher mean Sequential Organ Failure Assessment scores at the time of HFNC initiation (7.21 vs. 4.27, $p < 0.01$). The failure group also had higher in-hospital all-cause mortality, longer total duration of HFNC therapy, and longer hospital length of stay than the nonfailure group. Most HFNC failures (285/317 patients, 90.0%) occurred due to intubation, whereas 32 of 317 failures (10.0%) occurred due to death. Compared with patients who failed due to death, those who failed due to intubation were younger (median [IQR] 62 [53–72] vs. 87 [80–89], $p < 0.01$) and had significantly higher proportion of Black patients ($n = 153$ [53.7%] vs. $n = 10$ [31.3%], $p = 0.02$).

Model Performances for Predicting HFNC Failure

XGB model had the highest AUROC for predicting HFNC failure (0.707; 95% CI, 0.650–0.765), followed by the LR model (0.673 [0.612–0.735]), the SVM model (0.657 [0.597–0.717]), the ROX index (0.616 [0.546–0.685]), and the KNN model (0.526 [0.461–0.592]) (**Fig. 1**). Only the XGB model had significantly higher AUROC than the ROX index ($p = 0.01$); AUROC of the LR, SVM, and KNN models were not significantly different from that of the ROX index ($p = 0.16$, 0.31, 0.06, respectively). Other performance metrics for the XGB model and the ROX index based on the optimal cutoff values are presented in **eFigure 1** and **eTable 5** (http://links.lww.com/CCX/B319).

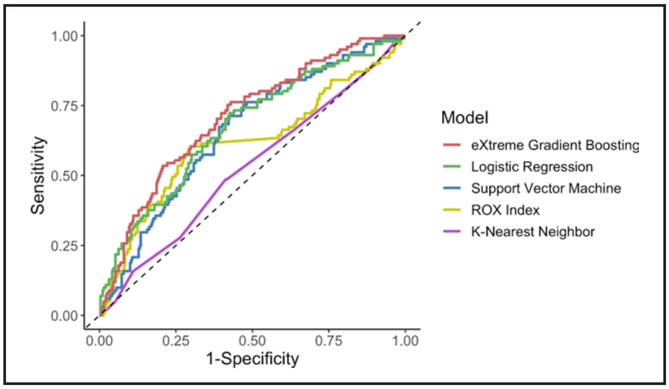

Figure 1. Receiver-operator characteristic curves of machine learning models and the respiratory rate-oxygenation (ROX) index for predicting high-flow nasal cannula failure.

Important Predictors for HFNC Failure

SHAP plot was generated to interpret the XGB model (the best performer) and understand the variable contributions to the model predictions for HFNC failure (**Fig. 2**). The two most important predictor variables for HFNC failure were minimum arterial Pao_2 and minimum arterial pH from ABG data. The remainder of top 20 most important variables mostly consisted of statistical measures of vital signs and common laboratory values, as well as older age and norepinephrine requirements contributing significantly to the model predictions for HFNC failure.

Stratified Analyses

Predictions for HFNC failure were stratified by race and compared between Black and White patients. The XGB model had significantly lower AUROC in Black patients than in White patients (AUROC [95% CI] 0.663 [0.586–0.740] vs. 0.808 [0.717–0.900], respectively, $p = 0.02$) (**Fig. 3**). All other models and the ROX index demonstrated nonsignificantly lower AUROC in Black patients compared with White patients (**eTable 6**, http://links.lww.com/CCX/B319). Stratification of the XGB model by sex and age group (< 65 vs. ≥ 65 yr old) demonstrated similar model performance between the subgroups (**eTable 7**, http://links.lww.com/CCX/B319).

Model Calibration and Sensitivity Analyses

To evaluate the racial differences in the accuracy of the XGB model, the XGB model calibration was assessed by generating calibration belts (32) (**eFig. 2**, http://links.lww.com/CCX/B319). The model was slightly miscalibrated in the overall cohort, which appeared to be driven by significant model miscalibration in Black patients that resulted in both over- and underestimation of risk of failure among Black patients. The calibration belt for White patients indicated that the model was well-calibrated and did not significantly overestimate or underestimate the risk of failure among White patients.

To investigate the potential causes of the racial differences in the accuracy of the XGB model, we first

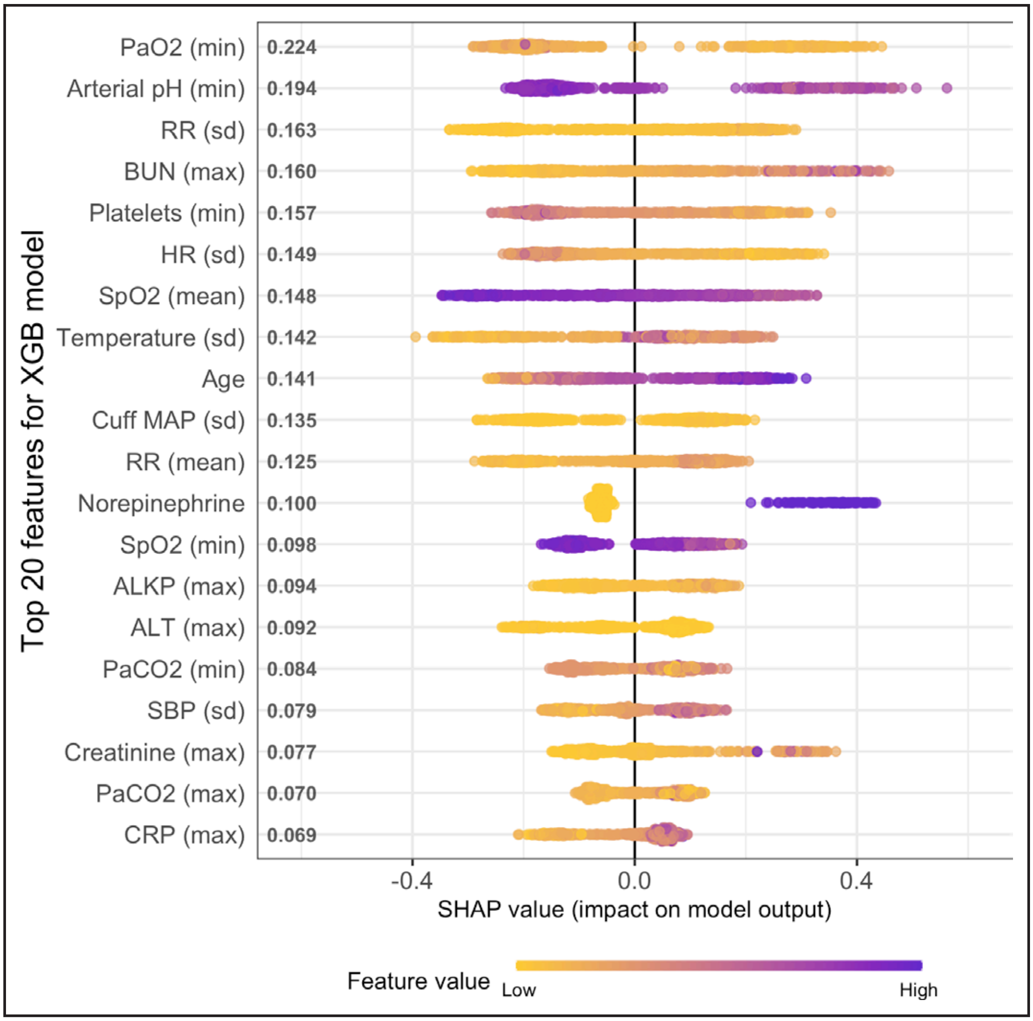

Figure 2. SHapley Additive exPlanations (SHAP) plot for the eXtreme Gradient Boosting (XGB) model for predicting high-flow nasal cannula (HFNC) failure. The top 20 predictors of HFNC failure in the XGB model are listed in the order of importance on the vertical axis. The distribution of color-coded data points on the horizontal axis indicates whether high (*purple*) versus low (*yellow*) values of that predictor are associated with positive SHAP value (contributing to prediction of failure) versus negative SHAP value (contributing to prediction of nonfailure). ALKP = alkaline phosphatase, ALT = alanine transferase, BUN = blood urea nitrogen, CRP = C-reactive protein, HR = heart rate, MAP = mean arterial pressure, max = maximum, min = minimum, RR = respiratory rate, SBP = systolic blood pressure.

compared the baseline characteristics and data distributions of important predictor variables (**eTable 8**, http://links.lww.com/CCX/B319). Black patients were started on HFNC earlier than White patients and had minor differences in a few predictor variables, but there were no significant differences in the rates of HFNC failure, NIV or IMV requirement, or missingness of ABG data between Black and White patients.

Then, we sequentially removed predictor variables with high potential for error in Black patients (e.g., Spo_2). The racial differences in accuracy of the model persisted even after sequentially removing Spo_2, respiratory rate, and all vital sign data from the XGB model inputs (**eTable 9**, http://links.lww.com/CCX/B319).

To investigate the impact of biased imputation, several strategies for handling missing data were tested. When the XGB model was developed and validated only on patients with nonmissing ABG data, there was no significant difference in accuracy between Black and White patients (AUROC [95% CI] 0.629 [0.525–0.733] vs. 0.669 [0.528–0.809], respectively, p = 0.65). Racial differences in model accuracy were still present when using PMM imputation or when missing data was left as missing, and there was a trade-off in which the strategy with the best overall model accuracy also had the worst racial bias (eTable 9, http://links.lww.com/CCX/B319).

Finally, to investigate biases in decision-making on intubation thresholds, models were developed and validated for predicting HFNC failure due to intubation only (rather than the original composite outcome). The models still had higher AUROC in White patients than in Black patients, although the difference between Black and White patients for the XGB model was no longer statistically significant (**eTable 10**, http://links.lww.com/CCX/B319). Additionally, an XGB model was developed and validated for predicting only mortality as the outcome (rather than the composite outcome). This model had similar accuracy for Black and

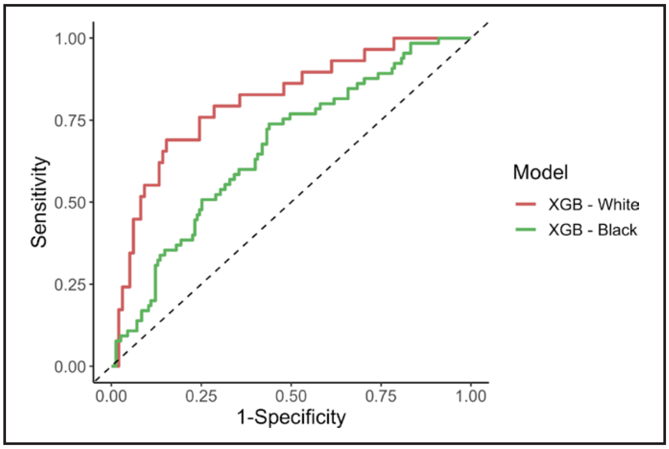

Figure 3. Receiver-operator characteristic curves of eXtreme Gradient Boosting (XGB) model for predicting high-flow nasal cannula failure, stratified by race.

White patients (AUROC [95% CI] 0.784 [0.695–0.873] vs. 0.749 [0.612–0.886], respectively, $p = 0.67$).

DISCUSSION

In this study, we developed and validated ML models to predict HFNC treatment failure in hospitalized COVID-19 patients, compared its performance to the ROX index, and evaluated potential racial differences in accuracy of the prediction models. The XGB model had the best performance and predicted HFNC failure in COVID-19 patients with better discrimination than the ROX index, using routinely available EMR data. However, the advantage of the XGB model over the ROX index may not be clinically substantial, and the XGB model performed significantly worse in Black patients compared with White patients, potentially due to inaccuracies in vital sign measurements, strategies for handling missing data, and differences in intubation thresholds. The findings suggest the need for intentional evaluation of performance across subgroups and an investigation into potential causes of biases in clinical prediction models.

The ROX index has been studied both in patients with pneumonia (20, 21) and COVID-19 (6–8, 22–26) with high risk of HFNC treatment failure. However, prior studies employed different thresholds of ROX index scores and inconsistent outcome definitions (with varying combinations of NIV, IMV, death, and/or weaning from HFNC used to define failure), and the precise clinical application of the ROX index remains unclear (33). Furthermore, the ROX index only reflects the patient's oxygenation and respirations (33), and incorporating additional parameters such as heart rate (34) or laboratory data (7) have yielded better predictive ability than the ROX index alone. Nonetheless, the ROX index performed well in our study with only one of four ML models (XGB) outperforming it, highlighting its strength as a parsimonious, easily calculable model that uses only three simple parameters to predict HFNC failure. Such predictive models have the potential to promptly identify high-risk patients for HFNC failure and allow closer monitoring, as well as to guide important decisions regarding early intubation and reduce the potential complications associated with delayed intubation in COVID-19 patients (35).

This study also highlighted that the added complexity of the ML models did not necessarily result in a clinically substantial improvement in the predictive performance of the models and that potential sources of bias will need to be better understood and mitigated before the ML models can be implemented. In particular, the XGB model was notable for performing significantly worse in Black patients compared with White patients in predicting HFNC failure. Two potential mechanisms of bias were considered (36). First, this may partially be due to accuracy-fairness trade-off or an artifact of the XGB algorithm, in which the hyperparameter tuning and learning process may have amplified bias by sacrificing fairness to make predictions as accurate as possible (36, 37). There is evidence of other clinical ML models that demonstrated increasing racial bias with increasing accuracy of the ML models, with the XGB model having the highest accuracy but also the most bias (38). Second, there may be inherent biases contained within our data that resulted in biased predictions (36), as the other models in our study also demonstrated small, nonsignificant differences in accuracy favoring White patients. Given the increasing concern that ML models being trained on data that reflects racial disparities in healthcare can propagate such disparities in their predictions (9, 15, 27), we sought to identify and mitigate the potential sources of biases.

First, we hypothesized that inaccuracies in pulse oximetry contributed at least partially to the racial differences in predictive performance of the XGB model, given that pulse oximetry has been consistently shown to be less accurate in Black patients (16–18), and even the ROX index also had slightly worse performance in

Black patients (although not statistically significant). However, removing pulse oximetry variables from the XGB model did not reduce the racial bias, suggesting that potential inaccuracies in pulse oximetry or vital signs did not completely account for the discrepancies.

Next, we also explored potential biases arising from missingness or strategies to handle missingness. When the XGB model was restricted to patients with nonmissing ABG data, the racial differences in the model accuracy were reduced. Given that missing data may represent biases in obtaining data and cannot be assumed to have the same distribution of true values for different subgroups, median imputation was likely an overly simplistic approach that provided more accurate estimates of missing ABG values for White patients than for Black patients. Notably, additional strategies for handling missing data did not significantly reduce the racial differences in model accuracy, and the strategy that improved the overall model accuracy was associated with worsened bias. Given that ABG variables were the two most important predictors in the XGB model, the handling of frequently missing ABG values likely impacted the racial bias in the XGB model, and highlights an important need to ensure that imputation strategies are thoughtfully designed to provide equitable assumptions.

Lastly, potential biases in intubation thresholds may have also contributed to inaccuracy. The decision to intubate is a complex decision dependent on numerous factors such as the patients' comorbidities, organ failures, preferences for life-sustaining treatment, as well as physicians' or institutional practices. Furthermore, evidence suggests that Black and Asian patients are less likely to receive IMV than White patients after meeting certain physiologic thresholds (19). Such differences in intubation thresholds, if present in our cohort, may have contributed to biased results when models were trained to predict a composite outcome that included physicians' decision to intubate. This hypothesis is supported by the finding that the models still demonstrated racial biases even when HFNC failure was defined only by intubation, but the XGB model had similar accuracy in Black and White patients when predicting mortality (thereby removing the potential impact of varying intubation thresholds in the outcome definition).

Our study has several strengths. We included a diverse cohort of patients from four different hospitals across a large metropolitan area. Our model uses routinely available clinical data to predict a clinically important outcome in patients with COVID-19, with improved accuracy over the ROX index. To our knowledge, our study is the first to investigate racial differences in the accuracy of predictive models for HFNC failure, which warrants further research. There are also several weaknesses. This was a retrospective analysis of EMR data from a single healthcare system, and prospective and/or external validation is necessary. Although our sensitivity analyses yielded some insights regarding potential causes of racial bias in our model, further investigation is needed to understand these causes and develop appropriate mitigation strategies. Lastly, this study only included patients with COVID-19, and similar ML models for predicting HFNC failure need to be evaluated in non-COVID-19 respiratory failure.

CONCLUSIONS

In conclusion, the novel ML model using XGB predicted HFNC failure in patients with COVID-19 with higher accuracy than the ROX index. However, the ML model performed worse in Black patients compared with White patients. Potential biases in measurements, missing data and imputation strategies, and intubation thresholds may have contributed to such bias in our model. Importantly, our study emphasizes the need for all new ML models to be evaluated for accuracy by race, and the need to investigate potential causes of biases. Future studies should also investigate mitigation strategies to ensure equitable performance and application of clinical ML models for patients of diverse backgrounds.

[1] *Division of Pulmonary, Allergy, Critical Care, and Sleep Medicine, Emory University, Atlanta, GA.*

[2] *Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA.*

[3] *Department of Surgery, Duke University School of Medicine, Durham, NC.*

[4] *Department of Radiology and Imaging Sciences, Emory University School of Medicine, Atlanta, GA.*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccejournal).

Drs. Yang, Gregory, and Bhavani were involved in conceptualization. Drs. Yang, Gregory, Robichaux, and Bhavani were involved *in data curation. Drs. Yang and Bhavani were involved in formal analysis, investigation, and methodology. Dr. Robichaux was involved in resources. Drs. Yang, Gregory, and Bhavani were involved in writing–original draft. Drs. Yang, Gregory, Gichoya, Holder, Martin, Esper, Kamaleswaran, and Bhavani were involved in writing–review and editing. All authors read and approved the final article.*

Dr. Holder is supported by the National Institute of General Medical Sciences of the National Institutes of Health (NIGMS/NIH) under award numbers K23GM137182 and 3K23GM137182-03S1. Dr. Kamaleswaran is supported by the NIGMS/NIH under award numbers R01GM139967 and 3R01GM139967-02S1. Dr. Gichoya is supported by a 2022 Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program and declares support from RSNA Health Disparities grant (number EIHD2204), Lacuna Fund (number 67), Gordon and Betty Moore Foundation, and NIH (NIBIB) MIDRC grant under contracts 75N92020C00008 and 75N92020C00021. Dr. Bhavani is supported by the NIGMS/NIH under award number K23GM144867. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The remaining authors have disclosed that they do not have any potential conflicts of interest.

For information regarding this article, E-mail: philip.yang@emory.edu

REFERENCES

1. Crimi C, Pierucci P, Renda T, et al: High-flow nasal cannula and COVID-19: A clinical review. *Respir Care* 2022; 67:227–240
2. Ospina-Tascon GA, Calderon-Tapia LE, Garcia AF, et al; HiFLo-Covid Investigators: Effect of high-flow oxygen therapy vs conventional oxygen therapy on invasive mechanical ventilation and clinical recovery in patients with severe COVID-19: A randomized clinical trial. *JAMA* 2021; 326:2161–2171
3. Mellado-Artigas R, Ferreyro BL, Angriman F, et al; COVID-19 Spanish ICU Network: High-flow nasal oxygen in patients with COVID-19-associated acute respiratory failure. *Crit Care* 2021; 25:58
4. Gershengorn HB, Pavlov I, Perez Y, et al: High-flow nasal cannula failure odds is largely independent of duration of use in COVID-19. *Am J Respir Crit Care Med* 2022; 205:1240–1243
5. Garner O, Dongarwar D, Salihu HM, et al: Predictors of failure of high flow nasal cannula failure in acute hypoxemic respiratory failure due to COVID-19. *Respir Med* 2021; 185:106474
6. Chandel A, Patolia S, Brown AW, et al: High-flow nasal cannula therapy in COVID-19: Using the ROX index to predict success. *Respir Care* 2021; 66:909–919
7. Xu J, Yang X, Huang C, et al: A novel risk-stratification models of the high-flow nasal cannula therapy in COVID-19 patients with hypoxemic respiratory failure. *Front Med (Lausanne)* 2020; 7:607821
8. Prakash J, Bhattacharya PK, Yadav AK, et al: ROX index as a good predictor of high flow nasal cannula failure in COVID-19 patients with acute hypoxemic respiratory failure: A systematic review and meta-analysis. *J Crit Care* 2021; 66:102–108
9. Allen A, Mataraso S, Siefkas A, et al: A racially unbiased, machine learning approach to prediction of mortality: Algorithm development study. *JMIR Public Health Surveill* 2020; 6:e22400
10. Ho LO, Li H, Shahidah N, et al: Poor performance of the modified early warning score for predicting mortality in critically ill patients presenting to an emergency department. *World J Emerg Med* 2013; 4:273–278
11. Ashana DC, Anesi GL, Liu VX, et al: Equitably allocating resources during crises: Racial differences in mortality prediction models. *Am J Respir Crit Care Med* 2021; 204:178–186
12. Hao B, Hu Y, Sotudian S, et al: Development and validation of predictive models for COVID-19 outcomes in a safety-net hospital population. *J Am Med Inform Assoc* 2022; 29:1253–1262
13. Hong C, Pencina MJ, Wojdyla DM, et al: Predictive accuracy of stroke risk prediction models across black and white race, sex, and age groups. *JAMA* 2023; 329:306–317
14. Huang J, Galal G, Etemadi M, et al: Evaluation and mitigation of racial bias in clinical machine learning models: Scoping review. *JMIR Med Inform* 2022; 10:e36388
15. Obermeyer Z, Powers B, Vogeli C, et al: Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366:447–453
16. Jubran A, Tobin MJ: Reliability of pulse oximetry in titrating supplemental oxygen therapy in ventilator-dependent patients. *Chest* 1990; 97:1420–1425
17. Sjoding MW, Dickson RP, Iwashyna TJ, et al: Racial bias in pulse oximetry measurement. *N Engl J Med* 2020; 383:2477–2478
18. Seitz KP, Wang L, Casey JD, et al: Pulse oximetry and race in critically ill adults. *Crit Care Explor* 2022; 4:e0758
19. Yarnell CJ, Johnson A, Dam T, et al: Do thresholds for invasive ventilation in hypoxemic respiratory failure exist? A cohort study. *Am J Respir Crit Care Med* 2023; 207:271–282
20. Roca O, Messika J, Caralt B, et al: Predicting success of high-flow nasal cannula in pneumonia patients with hypoxemic respiratory failure: The utility of the ROX index. *J Crit Care* 2016; 35:200–205
21. Roca O, Caralt B, Messika J, et al: An index combining respiratory rate and oxygenation to predict outcome of nasal high-flow therapy. *Am J Respir Crit Care Med* 2019; 199:1368–1376
22. Zhou X, Liu J, Pan J, et al: The ROX index as a predictor of high-flow nasal cannula outcome in pneumonia patients with acute hypoxemic respiratory failure: A systematic review and meta-analysis. *BMC Pulm Med* 2022; 22:121
23. Blez D, Soulier A, Bonnet F, et al: Monitoring of high-flow nasal cannula for SARS-CoV-2 severe pneumonia: Less is more, better look at respiratory rate. *Intensive Care Med* 2020; 46:2094–2095
24. Calligaro GL, Lalla U, Audley G, et al: The utility of high-flow nasal oxygen for severe COVID-19 pneumonia in a resource-constrained setting: A multi-centre prospective observational study. *EClinicalMedicine* 2020; 28:100570
25. Prower E, Grant D, Bisquera A, et al: The ROX index has greater predictive validity than NEWS2 for deterioration in COVID-19. *EClinicalMedicine* 2021; 35:100828
26. Abe T, Takagi T, Fujii T: Update on the management of acute respiratory failure using non-invasive ventilation and pulse oximetry. *Crit Care* 2023; 27:92

27. Vyas DA, Eisenstein LG, Jones DS: Eisenstein LG and Jones DS: Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *N Engl J Med* 2020; 383:874–882

28. Chen Y, Lin HY, Tseng TS, et al: Racial differences in data quality and completeness: Spinal cord injury model systems' experiences. *Top Spinal Cord Inj Rehabil* 2018; 24:110–120

29. Fongwa MN, Cunningham W, Weech-Maldonado R, et al: Comparison of data quality for reports and ratings of ambulatory care by African American and White Medicare managed care enrollees. *J Aging Health* 2006; 18:707–721

30. Moons KG, Altman DG, Reitsma JB, et al: Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. *Ann Intern Med* 2015; 162:W1–73

31. Leisman DE, Harhay MO, Lederer DJ, et al: Development and reporting of prediction models: Guidance for authors from editors of respiratory, sleep, and critical care journals. *Crit Care Med* 2020; 48:623–633

32. Nattino G, Lemeshow S, Phillips G, et al: Assessing the calibration of dichotomous outcome models with the calibration belt. *Stata J* 2017; 17:1003–1014

33. Junhai Z, Jing Y, Beibei C, et al: The value of ROX index in predicting the outcome of high flow nasal cannula: A systematic review and meta-analysis. *Respir Res* 2022; 23:33

34. Goh KJ, Chai HZ, Ong TH, et al: Early prediction of high flow nasal cannula therapy outcomes using a modified ROX index incorporating heart rate. *J Intensive Care* 2020; 8:41

35. Riera J, Barbeta E, Tormos A, et al; CIBERESUCICOVID Consortium: Effects of intubation timing in patients with COVID-19 throughout the four waves of the pandemic: A matched analysis. *Eur Respir J* 2023; 61:2201426

36. Liu S, Vicente LN: Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *Comput Manage Sci* 2022; 19:513–537

37. Tizpaz-Niari S, Kumar A, Tan G, et al: Fairness-aware configuration of machine learning libraries. Proceedings of the 44th International Conference on Software Engineering 2022;909–920.

38. Barton M, Hamza M, Guevel B: Racial equity in healthcare machine learning: Illustrating bias in models with minimal bias mitigation. *Cureus* 2023; 15:e35037