Mathematica
Progress Together

Health Research Brief

Authors: Emma Pendl-Robinson, Camille Shao, and Jennifer Starling, on behalf of N3C consortium

Mitigating Bias to Improve Fairness in Predictive Risk Modeling Using Healthcare Data: An Analysis of Long COVID Risk

Background. Algorithmic bias in healthcare predictive risk models can worsen existing health inequities, making bias mitigation crucial for responsible model development and implementation. Our study examined ways to improve fairness across both univariable and multivariable protected attributes using leading bias mitigation methods and measures of performance and fairness, aiming to provide researchers with guidance for how to test and improve algorithmic fairness. We conduct our analysis using predictive risk models for long COVID, an area of significant societal interest, as a case study to demonstrate effective strategies for addressing bias in predictive modeling.

Data sources. Our study used previously developed long COVID machine learning models applied to a sample of 1.23 million participants from the National COVID Cohort Collaborative (N3C), a longitudinal EHR data repository from 80 sites in the United States with more than 8 million COVID-19 patients.

Methods. We analyzed model fairness for the protected attributes of sex, race, and ethnicity by comparing performance and fairness metrics before and after applying bias mitigation techniques. Our evaluation focused on three leading algorithmic bias mitigation methods: reweighting, MAAT (Mitigating Algorithmic Bias with Adversarial Training), and FairMask. The analysis included both single and multiple protected attributes, using performance metrics (AUROC [area under the receiver operating characteristic curve] and PRAUC [area under the precision-recall curve]) and common fairness metrics (equal opportunity, predictivity equality, and disparate impact).

Findings. Our results demonstrate that applying bias mitigation techniques can improve fairness while maintaining model performance, as observed through monitoring key performance and fairness metrics. Across a variety of bias mitigation techniques, FairMask achieved the most significant gains in fairness for single protected attribute, with minor trade-offs for other attributes. Reweighting was more effective at boosting predictive performance metrics, but when optimizing performance or fairness with respect to one specific protected attribute, the performance and fairness for other attributes varied.

Conclusion. When building predictive risk models in healthcare, researchers should carefully consider the inclusion of protected attributes, monitor key performance and fairness metrics, and implement strategies for mitigating bias where needed. Testing and improving algorithmic fairness will ensure that predictive models contribute to more equitable healthcare outcomes, where algorithmic findings – such as those from long COVID predictive risk models – directly influence patient care, clinical decision-making, and policy.

Introduction

Model fairness

Machine learning models have gained popularity across various fields, including health services research, due to their ability to predict outcomes for unseen data.[1] These models can be used to help us make decisions because of their ability to synthesize large amounts of input data to make predictions. However, a model's predictive performance might be higher or lower for various subpopulations due to biases within the training data or model specification. Unchecked predictive models can perpetuate disparities and embed bias into systems, potentially harming certain groups.

With the rise in use of machine learning models, researchers are thinking about how to measure algorithmic fairness and mitigate algorithmic bias in predictive models that use health care data, such as EHRs.[2] In this white paper, we explore various methods of measuring algorithmic fairness and mitigating bias in predictive risk models built using EHR data. We provide results and examples in the context of identifying patients at risk of long COVID.

Background in long COVID

By November 2022, 94% of the US population was estimated to have contracted COVID at least once.[3]This proportion continues to rise as COVID-19 remains a widespread public health threat. Some people recover quickly after the initial infection, while others (between 10% to 30%) continue to experience symptoms even months after the initial infection.[4] Long COVID is defined by the World Health Organization (2021) as persistent COVID symptoms and/or long-term complications following a probable or confirmed infection.[5]

We focused on models that predict patients' risk of developing long COVID, as many challenges exist around identifying patients with elevated risk (e.g. long-term effects may exhibit differently for different people, and difficulties in tracking long-term and varying symptoms). Predicting risk of long COVID can help the public health sector identify population segments at elevated risk for long COVID in order to design and implement targeted intervention strategies and help the healthcare sector prepare to understand the demand for long COVID treatment and provide better patient care.

Potential for bias in long COVID predictive modeling

Long COVID prediction is important because timely risk assessment of long COVID outcomes can improve patient care and inform policy and health care resource allocation. However, there are serious concerns about bias in long COVID identification and modeling.[6,7] Disparities by sex, race, and ethnicity, and social economic status in COVID-19 patient outcomes have been well-documented.[8,9] For example, while males have higher risk for severe COVID outcomes such as death and intensive care unit admission, females are more likely to suffer from long COVID.[10] In addition, compared with White patients, patients from racial and ethnic minority groups had significantly different odds of developing long COVID-related conditions and symptoms.[11]

Researchers must carefully examine model fairness across different subgroups to achieve optimal and fair clinical decision-making. In our literature review, we found recommendations on about how to avoid algorithmic bias, measure model fairness, and mitigate bias for predictive COVID-19 models, as well as for other models that use EHR data.[2,12] We found a study that predicted COVID-19 patient outcomes that checked if the models achieved similar area under the receiver operating characteristic curve (AUROC) scores for sex and race subgroups.[13] The authors concluded that the models were fair because the AUROC for all the subgroups were similar and above 80 percent, so the models could be equitably applied across these demographic characteristics—which aligns with federal guidelines that recognize 80 percent as a standard for evaluating fairness. However, we did not find peer-reviewed studies on predictive long COVID models that included checks for algorithmic fairness or bias mitigation.[6,14,15]

Our contribution

In this white paper, we use our previously developed long COVID risk prediction models as a starting point to explore algorithmic fairness and performance metrics and bias mitigation methods.[16] We optimize for protecting single attributes separately and multiple protected attributes simultaneously using various bias mitigation methods. Lastly, we make recommendations on the best metrics and methods to use for mitigating bias.

This paper provides guidance to other researchers for how to test and improve algorithmic fairness. Ensuring the fairness and equity of models is an essential consideration in healthcare settings, where algorithmic findings – such as those from long COVID predictive risk models – directly influence patient care, clinical decision-making, and policy.

Methods

Data source and model specification

We used the same analytic sample, covariates, model specifications, and bootstrapping methods as our previously developed long COVID predictive risk models.[16] The sample included 1.23 million participants from the National COVID Cohort Collaborative (N3C), a longitudinal EHR data repository with information on more than 8 million COVID-19 patients from 80 sites in the United States. We defined long COVID using three symptom clusters: fatigue symptoms (4.7 percent of patients in sample), respiratory symptoms (2.5 percent of patients in sample), and cognitive symptoms (0.2 percent of patients in sample) (Table A.1).

We trained two types of machine learning models—binary logistic regression (LR) and binary random forest (RF)—on all three long COVID symptom clusters separately. We used a 70-30 train-test split; because the symptom clusters were rare outcomes, we trained the models on downsampled data and tested them on the original unbalanced test data. Results are averaged across 100 bootstrap samples from the training data and for the fatigue symptom cluster models, unless otherwise specified. We provide results for the cognitive and respiratory symptom cluster models in the appendix.

Comparing fairness metrics

We evaluated the fairness of the model predictions using three fairness metrics: equal opportunity ratio (EOR), predictivity equality ratio (PER), and disparate impact ratio (DIR). All three metrics compare aspects of model performance between a privileged group and an unprivileged group. To model fairness for a single protected attribute, we considered our protected attributes as binary categories. We specified the unprivileged group as the second largest non-missing category for the demographic categories sex (male), race (Black or African American), and ethnicity (Hispanic or Latino). We specified all other demographic sex, race, and ethnicity categories as the privilege group. Because the U.S. federal government considers a fairness rating below 80 percent to constitute disparate impact, we aimed for our models to achieve fairness of near or above 80 percent.[17]

The DIR, also known as the statistical parity ratio, is the ratio of the percent predicted as positive for the privileged versus unprivileged groups. The DIR is useful when there is concern that historic bias could cause higher rates of mislabeled results for some subgroups than others in the training data. Uneven mislabeling rates could lead to inaccurate observed positive and negative values.[18] For example, there is potential bias in the patients included in the EHR data and in the rates that long COVID symptoms are correctly identified.

The EOR is the ratio of true positive rates for the unprivileged compared to the privileged group.[19] The EOR is helpful when there are concerns about higher false negative rates for the unprivileged group, which could be an equity concern if the model is used to identify patients for beneficial services. For example, this could be an issue if the long COVID models were used to identify which patients at high risk of long COVID could receive selective preventive long COVID treatment.

The PER is the ratio of false positives for the privileged group when compared with the unprivileged group. The PER is helpful when there are concerns about higher false positive rates in the unprivileged group, which could be an equity concern if the model is used to identify patients for punitive actions. For example, this might be an issue if an insurance company used long COVID models to identify patients at high risk of long COVID for the purpose of increasing their health insurance premium.

Comparing metrics of model performance

We checked that our models achieved similar performance across different subgroups by sex (male, female, and other), race (White, Black or African American, Asian, Native Hawaiian and Pacific Islander [NHPI], other, and missing/unknown), and ethnicity (Hispanic or Latino, not Hispanic or Latino, and missing/unknown).

We calculated model performance using two common metrics for binary classification models: AUROC and the area under the precision-recall curve (PRAUC). Both AUROC and PRAUC scores range from 0 to 1, with higher scores indicating stronger model performance. The two scores cannot be directly compared because they measure distinct aspects of model performance. Although PRAUC scores tend to be lower than AUROC scores, this does not mean that AUROC scores are better for measuring model performance.[20]

The AUROC measures a model's ability to distinguish between the positive and negative classes by computing the likelihood that the model makes an incorrect prediction (false positive or false negative). The AUROC weights all incorrect predictions equally and is robust against imbalanced data sets.[20]

The PRAUC measures a model's precision (the number of positives correctly identified compared with predicted positives) relative to the model's recall (the number of positives correctly identified compared with all positives). This measure focuses more on the positive class than the negative class, which helps evaluate the trade-off between reducing false positives and false negatives. However, when there is class imbalance, PRAUC could be biased toward subpopulations with a higher prevalence of positive predictions.[20]

Bias mitigation methods

If the performance and fairness metrics uncover bias between subgroups, various techniques can be applied to improve the model's fairness.

We selected three methods to mitigate algorithmic bias: reweighting, Mitigating Algorithmic Bias with Adversarial Training (MAAT), and FairMask. This decision was guided by a benchmarking paper, “Fairness Improvement with Multiple Protected Attributes,” which evaluated 11 methods of correcting fairness.[21] The paper primarily focuses on sex and race as protected attributes, but we were guided by its conclusions on multidimensional fairness and performance trade-offs. The benchmark has also been conducted on two health care-related data sets, making it relevant to our work.

Reweighting adjusts the training data (a “pre-processing” approach to bias reduction) by assigning different weights to different observations based on their protected attribute values and observed outcomes. This method reduces bias by giving higher weights to underrepresented or disadvantaged groups and lower weights to privileged groups. In this process, the training data is reweighted to balance the impact of protected groups, enabling the model to make fairer predictions for each subgroup. This approach ensures that the model does not disproportionately favor one group over another, particularly when optimizing for a single protected group such as sex or race.[22]

Mitigating Algorithmic Bias with Adversarial Training (MAAT) is an ensemble approach aimed at addressing the fairness-performance trade-off in machine learning. It is an ”in-processing” technique, meaning it reduces bias by adjusting the model-fitting process itself, rather than ‘pre-processing’ methods that modify the data beforehand. Unlike traditional ensemble methods that optimize for a single objective, MAAT combines two models: one focused on performance and the other on fairness. The fairness model uses a debugging technique for the training data which corrects biases in the data by transforming it via a "We’re All Equal" worldview. This technique resamples the biased data, balancing the representation of privileged and unprivileged groups. MAAT then averages the predictions of both models to produce a final decision, enhancing fairness without significantly compromising performance.[23]

FairMask, another ”in-processing” approach, operates by learning a separate classifier for each protected attribute, such as sex or race. This classifier predicts artificial values for the protected attribute, masking the true values during training and testing. By using artificial values, FairMask ensures that the protected attribute does not overly influence the model’s predictions. This approach produces fairer outcomes by reducing bias in the protected attribute without directly altering the model’s architecture or performance metrics.[22]

Extending mitigation methods to multiple attributes to improve fairness

Our mitigation methods can be easily adapted to improve fairness across multiple protected attributes. Here, "multiple attributes" means considering the intersections of all protected attributes to form subgroups. Specifically, we had three binary protected attributes—sex (male), ethnicity (Hispanic or Latino), and race (Black or African American)—which allow for eight possible combinations.

Optimizing for multiple attributes involves considering all eight subgroups rather than optimizing for a particular group such as the male group. Reweighting calculates weights using combinations of protected attributes. MAAT trains multiple fairness models with the We’re All Equal worldview for each protected variable individually and combines them to adjust the final prediction. Lastly, FairMask generates artificial values for all protected attributes individually to build the final model.

After mitigating bias across combinations of protected attributes, we assessed fairness and performance using the AUROC and PRAUC metrics. In addition to our single attribute measures, we also included intersectional metrics to capture the differences across subgroups formed by combinations of protected attributes. These metrics help identify performance gaps by measuring the disparities between the maximum and minimum values across subgroups. Specifically, we introduced three intersectional metrics: intersectional disparate impact difference (IDID), intersectional equal opportunity difference (IEOD), and intersectional predictive equality difference (IPED).

However, computational complexity presents a challenge in the multiple attributes setting. Adding attributes increases computational complexity exponentially, which can make computation unmanageable. Reweighting simply adjusts weights for attribute combinations and thus scales more efficiently than MAAT and FairMask, especially when there are large numbers of protected attributes. MAAT and FairMask require more computational resources in the multiple attributes setting because they involve either training multiple fairness models or generating artificial values for each subgroup by training independent models.

Results

Baseline model performance and fairness

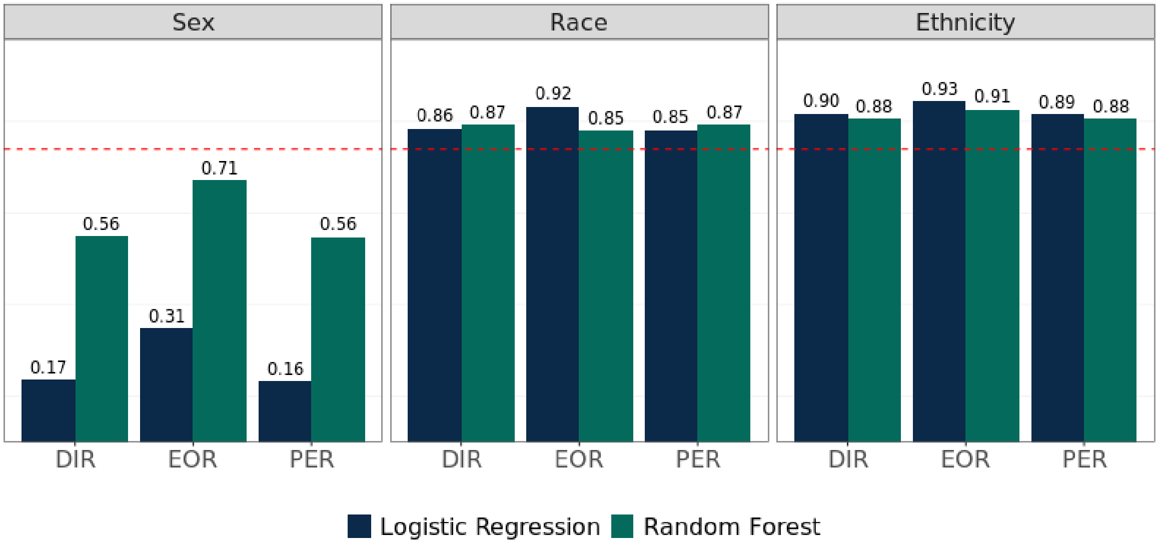
Model fairness

For sex, the RF model had noticeably higher fairness scores (DIR, EOR, PER) than the LR model. For the other protected attributes (race, ethnicity), the RF and LR models had similar fairness scores (Figure 1 and Table 1). This could indicate the RF model is a fairer machine learning model than LR before any bias mitigation method is applied.

The binary protected attribute race and ethnicity had high fairness scores at baseline for both the LR and RF models. However, the LR and RF models had lower fairness scores for sex. This indicates that the LR and RF models for race and ethnicity met the 0.80 disparate impact threshold but had meaningful differences in model predictions between the unprivileged (male) and privileged (female and other) sex subgroups.

For the sex protected attribute, EORs were higher than DIRs and PERs. This difference signals that there are smaller differences in the percent predicted as positive rates for patients in the privileged and unprivileged subgroups, and larger differences in the false positive rates between subgroups.

Figure 1. The models showed meaningful differences in fairness measures for the single protected attribute sex. However, the fairness metrics for race and ethnicity met the 80% threshold at baseline.



Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specification.

-- 80% fairness measure threshold

DIR = disparate impact ratio (also known as the statistical parity ratio); EOR = equal opportunity ratio; PER = predictive equality ratio.

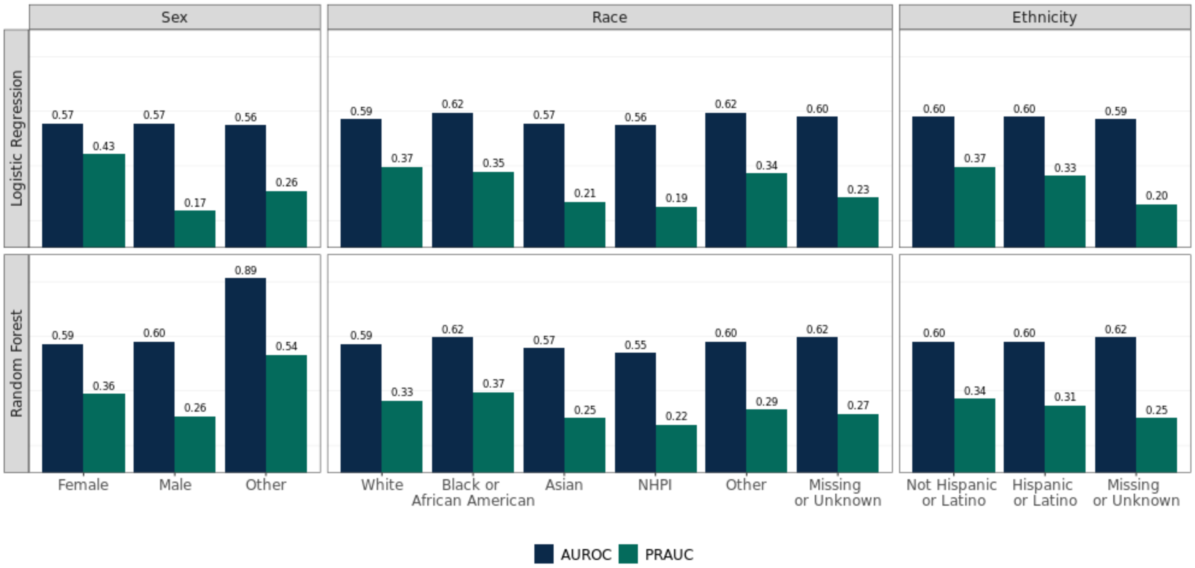
Model performance

Within the baseline LR and RF models, there were similar AUROC values for the sex, race, and ethnicity subgroups. This result indicates that our models had similar rates of correct predictions across subgroups (Figure 2 and Table 2). For the ethnic subgroups, the range of AUROCs was 1 percentage point for the LR model and 2 percentage points for the RF model. There were similarly small ranges of values for the sex subgroups (except for the "sex other" subgroup in the RF model). There was a slightly larger range of AUROC values for the racial subgroups, with all the AUROCs measuring around 0.60. The broad range is likely due to the small sample size of some racial subgroups, such as NHPI with fatigue long COVID symptoms (n = 61).

The one exception to the pattern of similar AUROCs within subgroups was the LR model "sex other" subgroup, which had a remarkably high AUROC. This was likely caused by the small number of patients in our sample who marked "sex other" *and* had the fatigue long COVID outcome (n < 20). For the single variable models, the sex subgroups "female" and "other" are combined; therefore, the small sample size for "sex other" had minimal effect on single variable model fairness.

The PRAUCs were lower than the AUROCs for all subgroups in both the RF and LR models. Whereas overall model accuracy is relatively high, the model precision (ratio of true positives to predicted positive) is low. Between the model subgroups, PRAUC values vary more than AUROC values, indicating there were higher rates of false positives for some subgroups than others. For example, for the RF sex subgroup, PRAUC for female (PRAUC = 0.36) is 10 percentage points higher than the male (PRAUC = 0.26).

Figure 2. At baseline, the models had mostly consistent AUROCs between subgroups, while PRAUCs were lower and had more variation between subgroups.



Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specification.

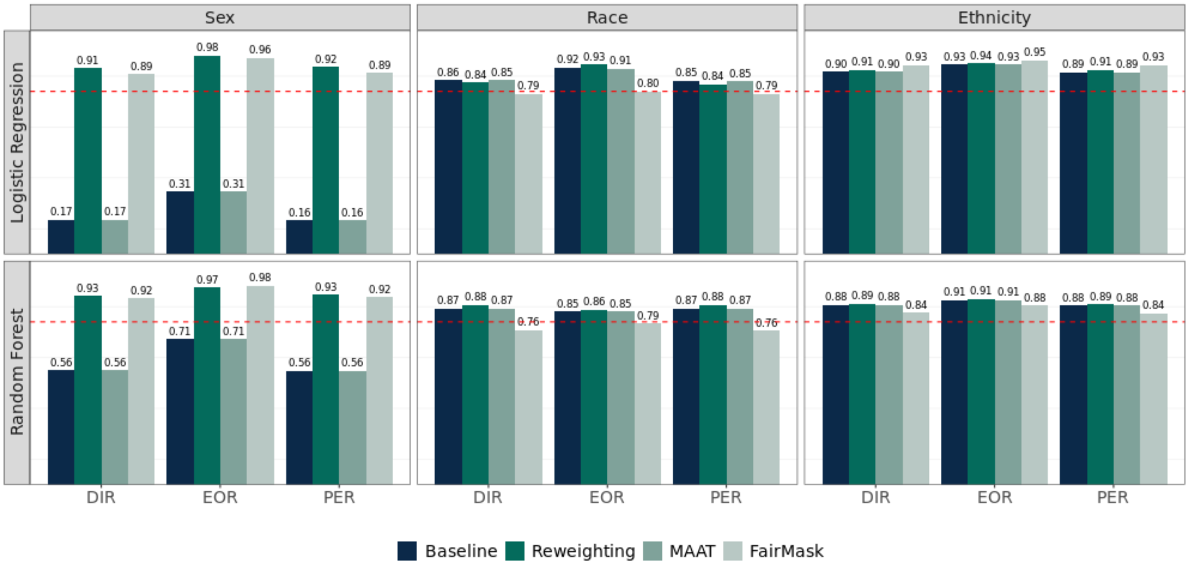
AUROC = area under the receiver operating characteristic curve; PRAUC = area under the precision-recall curve.

Mitigating bias for a single protected attribute

Model fairness

For the protected attribute sex, reweighting and FairMask showed substantial improvements in the fairness metrics for the LR model (Figure 3 and Table 1). MAAT had no effect on the fairness metrics for the LR model for the protected attribute sex. There was a similar pattern for the RF model. This indicates that reweighting and FairMask are preferable bias mitigation techniques for our LR and RF models.

Figure 3. When optimizing for the single protected attribute sex, reweighting and FairMask greatly improved the fairness measures, but MAAT did not. The bias mitigation methods did not meaningfully improve the fairness measures for the single protected attributes race or ethnicity.



Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

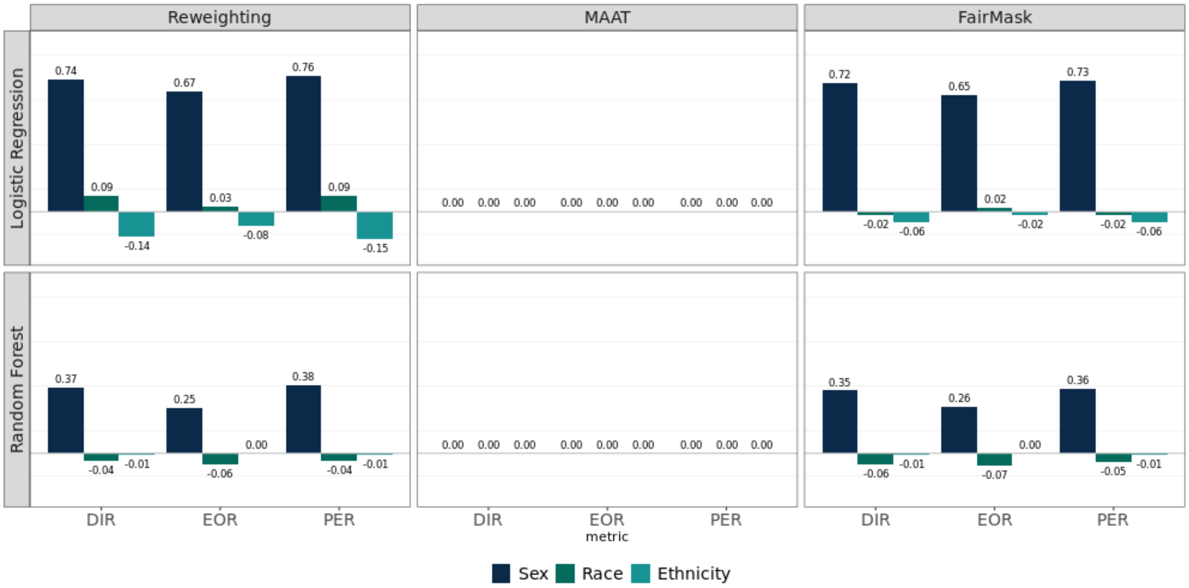
Note: Results are shown as the average across 100 bootstrap samples for each model specification. When optimizing bias mitigation for a single protected attribute, we specified the unprivileged group as the second largest non-missing category of the protected attribute: sex (male), race (Black or African American), and ethnicity (Hispanic or Latino).

-- 80% fairness measure threshold

DIR = disparate impact ratio (also known as the statistical parity ratio); EOR = equal opportunity ratio; MAAT = mitigating algorithmic bias with adversarial training; PER = predictive equality ratio.

However, mitigating bias for one protected attribute at a time can negatively impact the fairness of the other protected attributes. While we observed significant improvements when optimizing fairness for sex, a closer look reveals trade-offs in race and ethnicity. For example, using reweighting and FairMask to optimize sex fairness led to small decreases in fairness scores for race and ethnicity (Figure 4). Applying reweighting to the LR model led to decreased fairness (by 0.14 for DIR, 0.08 for EOR, and 0.15 for PER) for the ethnicity protected attribute. This was the largest decrease in fairness due to mitigating bias for a single protected variable. For most of the models, reweighting, MAAT, and FairMask had minimal effect on the fairness of the non-protected attributes.

Figure 4. Optimizing for the single protected attribute sex, improvements in fairness measures sometimes negatively affected fairness measures for race and ethnicity.



Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specification. When optimizing bias mitigation for the protected attribute sex, we specified the unprivileged group as male.

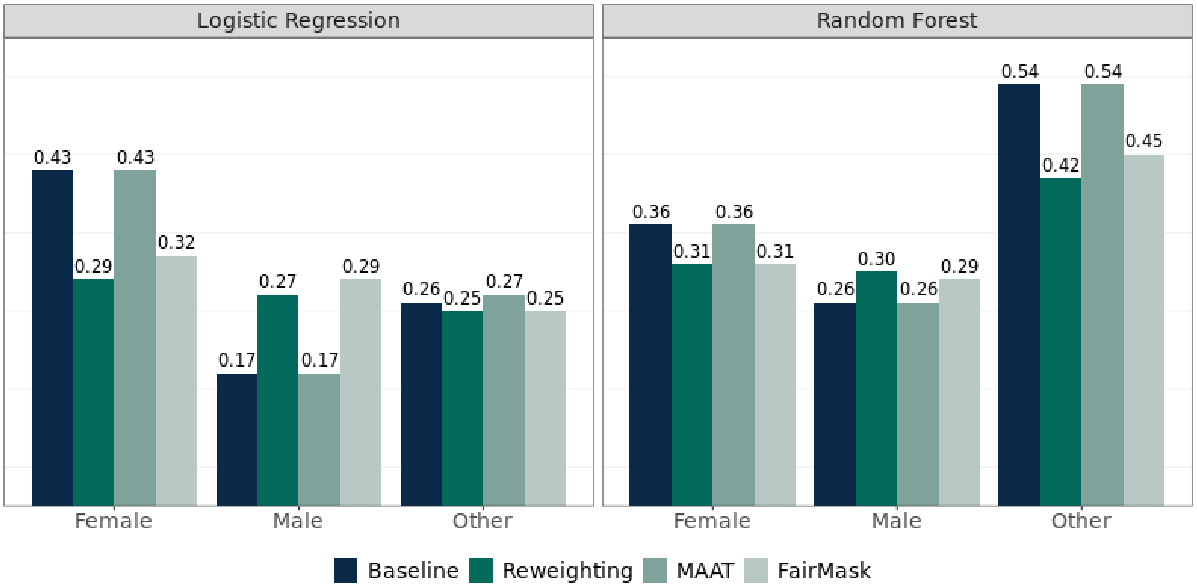
DIR = disparate impact ratio; EOR = equal opportunity ratio; PER = predictive equality ratio.

For the protected attributes race and ethnicity, none of the bias mitigation methods led to meaningful improvements in fairness metrics for either the LR or RF models (Figure 3). This could be because the fairness metrics values were high in the baseline models, so there was less room for improvement. Remarkably, FairMask decreased fairness metrics when compared with baseline for the protected variable race in the LR and RF models and ethnicity for the RF model. These results suggest that if a model's prediction at baseline is fair for the target protected attributes, mitigating bias might not change or might even decrease the model's fairness.

Model performance

For the protected attribute sex, reweighting and FairMask caused noticeable trade-offs between PRAUCs for the male and female subgroups for the fatigue LR and RF models (Figure 5 and Table 2). For the LR model, PRAUC for the male subgroup (the protected group) increased from baseline by 10 percentage points and by 12 percentage points with reweighting and FairMask. On the other hand, for the same model, PRAUC for the female subgroup decreased from baseline by 14 percentage points and by 11 percentage points with reweighting and FairMask, respectively. The RF models show similar directional changes in PRAUC for the male and female subgroups. Reweighting and FairMask brought PRAUC for the male and female subgroups closer together, which improved the fairness metrics. The AUROCs for male and female remained consistent or slightly improved with reweighting and FairMask. This indicates that the rates of false negatives and false positives changed proportionally, leading to the smaller visible effects when compared with PRAUC.

Figure 5. Reweighting and FairMask improved PRAUC for the male subgroup but decreased PRAUC for the female and sex other subgroups when optimizing for the single protected attribute sex.



Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specification. When optimizing bias mitigation for the protected attribute sex, we specified the unprivileged group as male.

MAAT = mitigating algorithmic bias with adversarial training; PRAUC = area under the precision-recall curve.

For the protected attributes of race and ethnicity, reweighting, MAAT, and FairMask had a less noticeable impact on the race and ethnicity subgroups' AUROC and PRAUC values for both the LR and RF models (Table 2). This is consistent with the smaller changes in fairness metrics for the racial and ethnic subgroups after mitigating bias.

Mitigating bias for multiple protected attributes

In this section, we present the results of optimizing all three protected attributes together. Although performance is shown for individual attributes like sex, the model was optimized for all attributes simultaneously, unlike earlier, where each attribute was optimized separately. We also introduce intersectional fairness metrics, where smaller values indicate reduced disparities and improved fairness across subgroups.

Fairness for multiple protected attributes

In sum, FairMask consistently delivered the most substantial improvements in fairness, particularly for the sex attribute, across LR and RF models. In LR, FairMask achieved significant gains for sex but lowered race fairness moderately. In RF, the trade-offs were more evenly distributed, resulting in smaller declines in fairness for race and minimal impact on ethnicity. Reweighting offered modest improvements in LR, especially for sex, but showed a slight decreasing trend across all attributes in RF. MAAT had little effect on fairness metrics in either model, remaining close to baseline (Figure 6 and Table 3).

Reweighting

Reweighting in LR showed moderate improvements for sex, indicating some success in enhancing fairness for this attribute. However, the changes for race and ethnicity were minimal, with fairness metrics remaining largely stable. This suggests that while reweighting offered a more balanced approach to fairness optimization when compared with FairMask, the overall impact was less substantial. The moderate gains in sex fairness did not come at a significant cost to race or ethnicity, making reweighting a less impactful but safer choice for fairness adjustments in LR.

In contrast, reweighting in RF resulted in very subtle changes across fairness metrics, with a slight decreasing trend for all attributes. For instance, DIR for race decreased by 0.05 for Black or African American and PER for ethnicity decreased by 0.03 for Hispanic or Latino. Although these changes were minimal, they highlight the lesser effectiveness of reweighting in RF when compared with LR, where it had a more positive impact on sex fairness. The minimal improvements suggest reweighting may be less suited to RF, where the baseline fairness was already higher.

Looking at intersectional metrics, reweighting presented more nuanced results. In LR, IEOD slightly decreased, reflecting improved fairness in true positive rates, while IDID and IPED increased slightly, indicating a rise in disparities for false positive rates and disparate impact. In RF, all intersectional metrics (IDID, IEOD, and IPED) increased, signaling a worsening of fairness. This could suggest reweighting may have over-prioritized optimizing the larger subgroups, worsening fairness for smaller groups in RF.

MAAT

MAAT showed no significant changes in fairness metrics for either LR or RF, with values remaining close to baseline across all protected attributes. This indicates MAAT's limited ability to mitigate biases when optimizing for multiple attributes. The lack of effect on intersectional metrics further reinforces MAAT's neutrality, as these values remained identical to baseline in both models.

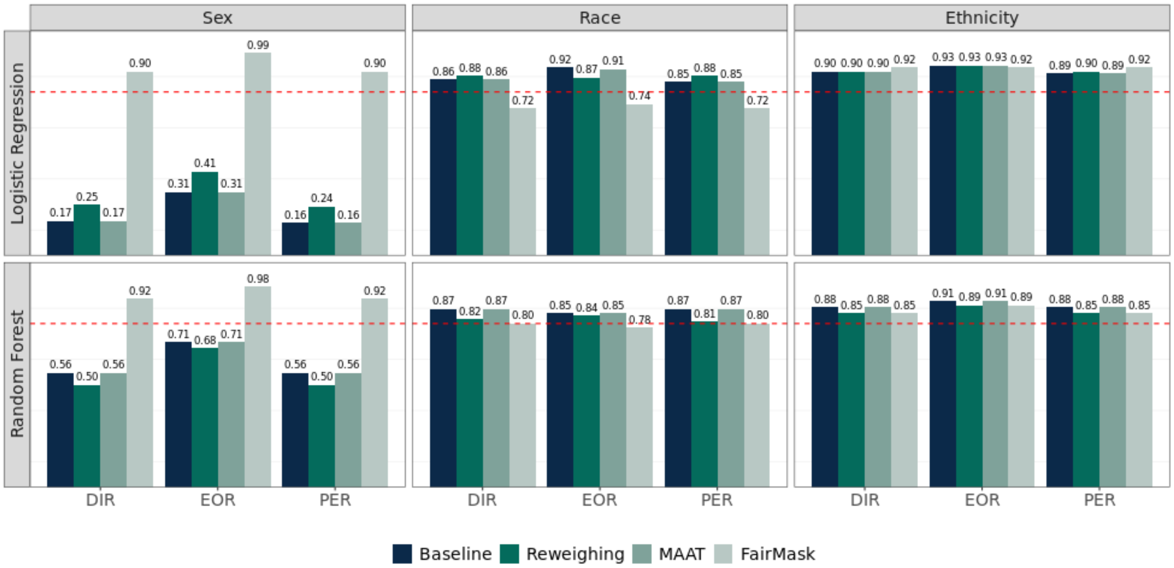
FairMask

In LR, FairMask achieved the most significant improvements for sex, reflecting the influence of the male subgroup, which comprised about 44 percent of the data set. These gains came with some reductions for race, indicating a higher cost to fairness for race in LR when compared with RF. Ethnicity remained largely stable. Despite these trade-offs, FairMask still managed to balance fairness across attributes, achieving significant improvements for sex without severely compromising fairness for race or ethnicity. Notably, fairness metrics for RF ended up higher than those for LR.

With RF, FairMask also significantly improved fairness for sex, but the trade-offs for race and ethnicity were more evenly distributed when compared with LR. The declines for race were modest and the impact on ethnicity was minimal. This indicates that FairMask in RF was more effective at balancing fairness improvements across all protected attributes, reducing disparities without significantly affecting any one attribute.

FairMask demonstrated the most significant reductions in intersectional fairness metrics across both LR and RF models. In LR, FairMask significantly reduced IDID, IEOD, and IPED, with similar reductions observed in RF. These improvements suggest FairMask reduced disparities between smaller subgroups and larger groups. Lower IEOD indicated that true positive rates for smaller subgroups, such as minorities, were more aligned with the majority group, while lower IPED shows that false positive rates were more comparable between the groups as well. Overall, FairMask substantially reduced gaps in prediction quality across demographic intersections, making it a fairer and more accurate method of optimization.

Figure 6. Optimizing multiple attributes together, FairMask improved fairness, especially for sex, with trade-offs in race for LR and more balanced effects in RF. Reweighting showed modest gains and MAAT had little impact.



Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specification. When mitigating bias, we specified the unprivileged group as the second largest non-missing category of the protected attribute: sex (male), race (Black or African American), and ethnicity (Hispanic or Latino).

-- 80% fairness measure threshold

DIR = disparate impact ratio (also known as the statistical parity ratio); EOR = equal opportunity ratio; MAAT = mitigating algorithmic bias with adversarial training; PER = predictive equality ratio.

Performance for multiple protected attributes

After applying bias mitigation techniques across all three protected attributes (sex, race, and ethnicity), the AUROC changes remained relatively small, typically within ±0.02. However, PRAUC values revealed more noticeable shifts. Reweighting generally showed better performance improvements, particularly in PRAUC, while FairMask had mixed impact. MAAT showed little to no change in performance metrics (Figure 7 and Table 4).

Reweighting

For LR, reweighting showed consistent improvements across most subgroups. PRAUC values increased for several subgroups, such as male, female, and white. The "sex other" subgroup, while representing a smaller portion of the population, saw minor improvement. For Hispanics or Latinos, the PRAUC increased moderately. Overall, reweighting had a positive effect on most subgroups, with only slight deviations in some cases. AUROC values, however, remained largely stable, with changes generally within ±0.02.

In the RF model, reweighting produced similarly positive results, with PRAUC improvements for the male subgroup, though the "sex other" subgroup experienced a slight decline. Despite this, the overall effect of reweighting was positive across subgroups. As with LR, AUROC changes remained small and close to baseline.

MAAT

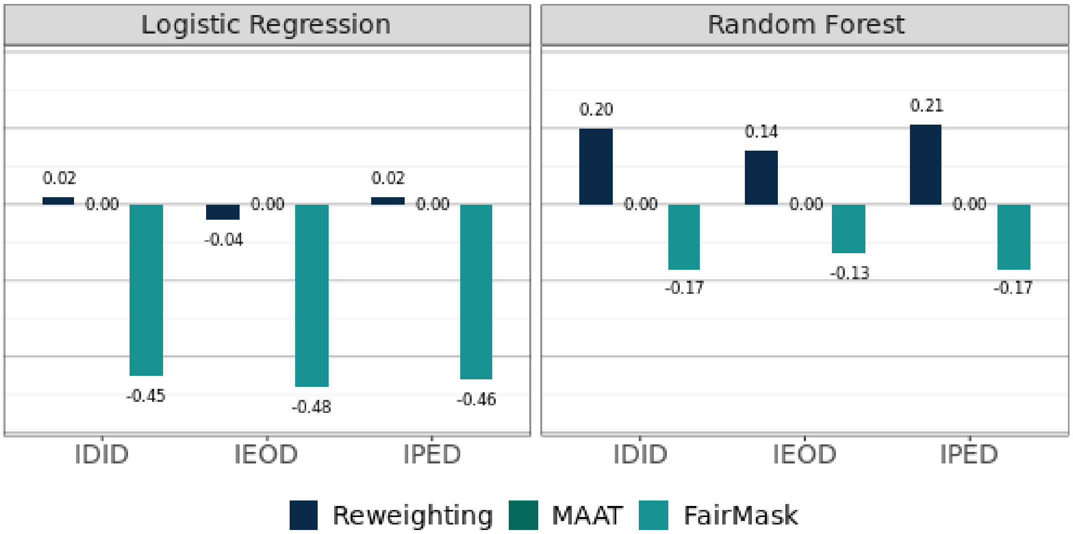
MAAT had a neutral impact on performance across all subgroups, in both LR and RF. Neither AUROC nor PRAUC exhibited significant changes, mirroring MAAT's minimal influence on fairness metrics. This stability suggests that MAAT had little effect on the models' performance metrics overall.

FairMask

FairMask demonstrated a more mixed impact on performance, especially in LR. Some subgroups, like female and Hispanic or Latino, saw declines in PRAUC, while the male subgroup experienced moderate improvement. While FairMask improved fairness metrics for the protected groups, it occasionally overcorrected, leading to performance declines in smaller subgroups. Nonetheless, AUROC remained stable within the ±0.02 range, indicating that ranking ability was less affected than precision-recall trade-offs.

In RF, FairMask's impact was similarly mixed, with subgroups like female and white experiencing declines, while male performance remained stable or improved slightly. As in LR, this suggests FairMask's overcorrection in smaller subgroups led to diminished performance.

Figure 7. Optimizing multiple attributes together, FairMask improved intersectional fairness by decreasing disparities across all models. Reweighting showed some gains in LR but worsened disparities in RF, and MAAT had no significant impact.



Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specification. When mitigating bias, we specified the unprivileged group as the second largest non-missing category of the protected attribute: sex (male), race (Black or African American), and ethnicity (Hispanic or Latino).

IDID= intersectional disparate impact difference; IEOD = intersectional equal opportunity difference; IPED = intersectional predictive equality difference; MAAT = mitigating algorithmic bias with adversarial training.

Discussion

Recommendations

We demonstrated application of the algorithmic fairness checks and bias mitigation methods to long COVID predictive models that use EHR data. Based on our analysis, we make the following recommendations to researchers developing predictive risk models using health care data.

Identify areas of potential bias. This includes identifying potential bias in research questions, data sources, and modeling techniques. Mathematica's blog post, Advancing Equity in Policy Research by Addressing Bias in Data Analytics, recommends additional steps to take at each step of the model development process.[1] For example, the Tackling Algorithmic Bias in Health Care blog post shows that bias can be mitigated proactively in the data cleaning and model selection processes.[24]

Select model type and parameterization with fairness in mind. We found that model selection and specification significantly influenced the baseline model's fairness. For example, in our analysis of the fatigue symptom cluster, the RF model was sometimes fairer than the LR model before applying bias mitigation methods. When developing a statistical model, researchers should consider how the model type and parameterization could affect the fairness of the results.

Select performance and fairness metrics based on the model's application. Intended use of predictive risk model outputs can inform the relative importance of various aspects of model performance. In other words, consider the potential negative effects of the model predicting an individual incorrectly or differences in model performance between subgroups.

Check for differences in model performance across demographic subgroups and intersectional subgroups at risk of bias. For our study, we aimed for the models to perform similarly among the demographic characteristics sex, race, and ethnicity. Depending on the model and research question, researchers might check model performance on other variables such as age, insurance status, or geographic location. Consider how subgroups intersect and how to address intersectional health equity. The blog post Intersectionality: Amplifying Impacts on Health Equity provides more guidance on intersectionality.[25]

Before mitigating bias, confirm that subgroup differences are caused by algorithmic bias, not inherent differences. Distinguish between subgroup difference caused by algorithmic bias versus inherent differences between subgroups. If protected groups exhibit natural variations in health outcomes due to biological, social, or environmental factors, it is important to assess whether observed disparities reflect true differences or biases introduced by the model. Mitigating bias without this understanding risks masking meaningful variation rather than addressing actual unfairness.

In our case, the improvements in fairness metrics (DIR, EOR, and PER) suggest some original disparities may have been due to bias in the model, but natural differences cannot be ruled out. Incorporating domain knowledge is crucial to determine if changes align with known health disparities.

Bias mitigation should aim to correct unjust disparities while acknowledging legitimate differences. Fairness metrics, combined with domain expertise, can help ensure that true biases are targeted without obscuring real-world patterns.

Understand the trade-offs presented by each bias mitigation method. When the baseline model fairness is insufficient, bias mitigation methods such as reweighting and FairMask can meaningfully improve model fairness without sacrificing model performance. In our case, FairMask provided the largest gains, especially for the sex attribute, with LR showing considerable improvements and RF achieving the best overall outcomes. On the other hand, reweighting offered reliable, steady improvements across subgroups without harming other attributes, making it a dependable option, even if the gains were smaller. This presents a choice between pursuing smaller but consistent gains across all attributes with reweighting, which is more computationally manageable, or opting for more substantial improvements in a single attribute such as sex using FairMask, with minor to moderate trade-offs in other variables and higher computational demands. Selecting a method of bias mitigation requires balancing fairness improvements with potential trade-offs based on the study's context, goals, and resources available.

Limitations

While our work aims to act as a proof of concept and provide general recommendations, the analysis has limitations. First, we limited the scope of analysis to two common performance metrics (AUROC and PRAUC), three common fairness metrics (equal opportunity, predictive equity, and statistical parity), and three promising bias mitigation techniques (reweighting, MAAT, and FairMask). Second, our fatigue symptom cluster long COVID models had low overall performance (AUROC: 0.60/0.60 and PRAUC: 0.36/0.37 for LR and RF, respectively). While these results align those of similar long COVID models, the models need to be improved before they are implemented.[14] Finally, we protected only the second largest subgroup within each protected attribute, which may not adequately address fairness for all underrepresented subgroups. In addition, the second largest subgroup may not correspond to a society-wide unprivileged group, as in our models, where the second largest sex subgroup was male. To ensure comprehensive fairness, it is important to consider protecting all subgroups within the protected category.[26]

Endnotes

[1]Newman A. Advancing equity in policy research by addressing bias in data analytics. Mathematica. July 10, 2024. Accessed September 18, 2024. https://www.mathematica.org/blogs/advancing-equity-in-policy-research-by-addressing-bias-in-data-analytics.

[2] Xu J, Xiao Y, Wang WH, et al. Algorithmic fairness in computational medicine. eBioMedicine. 2022;84. doi:10.1016/j.ebiom.2022.104250.

[3] Klaassen F, Chitwood M, Cohen T, et al. Changes in population immunity against infection and severe disease from SARS-CoV-2 Omicron variants in the United States between December 2021 and November 2022. *Clinical Infectious Diseases*. 2023; 77(3):355–361. https://doi.org/10.1093/cid/ciad210.

[4] Herman E, Shih E, and Cheng A. Long COVID: Rapid Evidence Review. *American Family Physician*. 2022;106(5):523–532.

[5] World Health Organization. A clinical case definition of post COVID-19 condition by a Delphi consensus. Published on October 6, 2021. Accessed November 25, 2024. https://www.who.int/publications/i/item/WHO-2019-nCoV-Post_COVID-19_condition-Clinical_case_definition-2021.1.

[6] Ahmad I, Amelio A, Merla A, Scozzari F. A survey on the role of artificial intelligence in managing long COVID. *Frontiers in Artificial Intelligence*. 2024;6. Accessed September 19, 2024. https://www.frontiersin.org/journals/artificial-intelligence/articles/10.3389/frai.2023.1292466.

[7] Pfaff ER, Madlock-Brown C, Baratta JM, et al. Coding long COVID: characterizing a new disease through an ICD-10 lens. *BMC Medicine*. 2023;21(1):58. doi:10.1186/s12916-023-02737-6.

[8] Kharroubi SA, Diab-El-Harake M. Sex-differences in COVID-19 diagnosis, risk factors and disease comorbidities: A large US-based cohort study. *Frontiers in Public Health*. 2022;10. doi:10.3389/fpubh.2022.1029190.

[9] Webb Hooper M, Nápoles AM, Pérez-Stable EJ. COVID-19 and racial/ethnic disparities. JAMA. 2020;323(24):2466-2467. doi:10.1001/jama.2020.8598.

[10] Cohen J, van der Meulen Rodgers Y. An intersectional analysis of long COVID prevalence. *International Journal for Equity in Health*. 2023;22(1):261. doi:10.1186/s12939-023-02072-5.

[11] Khullar D, Zhang Y, Zang C, et al. Racial/Ethnic Disparities in post-acute sequelae of SARS-CoV-2 infection in New York: an EHR-based cohort study from the RECOVER program. *Journal of General Internal Medicine*. 2023;38(5):1127-1136. doi:10.1007/s11606-022-07997-1.

[12] Tsai TC, Arik S, Jacobson BH, et al. Algorithmic fairness in pandemic forecasting: lessons from COVID-19. *npj Digital Medicine*. 2022;5(1):1-6. doi:10.1038/s41746-022-00602-z.

[13] Giuste FO, He L, Lais P, et al. Early and fair COVID-19 outcome risk assessment using robust feature selection. *Scientific Reports*. 2023;13(1):18981. doi:10.1038/s41598-023-36175-4.

[14] Antony B, Blau H, Casiraghi E, et al. Predictive models of long COVID. *eBioMedicine*. 2023;96. doi:10.1016/j.ebiom.2023.104777.

[15] Pfaff ER, Girvin AT, Bennett TD, et al. Identifying who has long COVID in the USA: a machine learning approach using N3C data. *The Lancet Digital Health*. 2022;4(7):e532-e541. doi:10.1016/S2589-7500(22)00048-6.

[16] Shanmugam P, Bair M, Pendl-Robinson E, Hu XC. Can longitudinal electronic health record data identify patients at higher risk of developing long COVID? Published online May 31, 2024. 2024.02.08.24302528. doi:10.1101/2024.02.08.24302528.

[17] 29 CFR § 1607.4. *29 CFR § 1607.4 - Information on Impact. - Content Details - CFR-2024-Title29-Vol4-Sec1607-4*. Vol 44 U.S.C. 3501et seq.; 2024. Accessed October 10, 2024. https://www.govinfo.gov/app/details/CFR-2024-title29-vol4/CFR-2024-title29-vol4-sec1607-4.

[18] Equality AI. Equality AI fairness metric selection questionnaire and tree. Published online December 22, 2022. Accessed April 19, 2024. https://github.com/EqualityAI/EqualityML/blob/main/Equality%20AI%20Fairness%20Metric%20Selection%20Questionnaire%20%26%20Tree.pdf.

[19] Bellamy R, Dey K, Hind M, et al. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. ArXiv. Published online October 3, 2018. Accessed June 13, 2024. https://www.semanticscholar.org/paper/AI-Fairness-360%3A-An-Extensible-Toolkit-for-and-Bias-Bellamy-Dey/c8541b1dc813f3a638d7acc79e5f972e77f3c5a7.

[20] McDermott MBA, Hansen LH, Zhang H, Angelotti G, Gallifant J. A closer look at AUROC and AUPRC under class imbalance. Published online April 18, 2024. doi:10.48550/arXiv.2401.06091.

[21] Chen Z, Zhang JM, Sarro F, Harman M. Fairness improvement with multiple protected attributes: how far are we? In: *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. ICSE '24. Association for Computing Machinery; 2024:1-13. doi:10.1145/3597503.3639083.

[22] Peng K, Chakraborty J, Menzies T. FairMask: Better fairness via model-based rebalancing of protected attributes. Published online October 27, 2022. doi:10.48550/arXiv.2110.01109.

[23] Chen Z, Zhang JM, Sarro F, Harman M. MAAT: a novel ensemble approach to addressing fairness and performance bugs for machine learning software. In: *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ESEC/FSE 2022. Association for Computing Machinery; 2022:1122-1134. doi:10.1145/3540250.3549093.

[24] Brown L, Whicher D, McGlone M. Tackling algorithmic bias in health care. Mathematica. November 10, 2021. Accessed October 7, 2024. https://www.mathematica.org/blogs/tackling-algorithmic-bias-in-health-care.

[25] Michaels E, Wesley DB, O'Neil S. Intersectionality: amplifying impacts on health equity. Mathematica. January 26, 2023. Accessed October 7, 2024. https://www.mathematica.org/blogs/intersectionality-amplifying-impacts-on-health-equity.

[26] Castelnovo A, Crupi R, Greco G, Regoli D, Penco IG, Cosentini AC. A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*. 2022;12:4209. doi:10.1038/s41598-022-07939-1.

Acknowledgements

This work is supported by Mathematica.

N3C attribution

The analyses described in this white paper were conducted with data or tools accessed through the NCATS N3C Data Enclave https://covid.cd2h.org and N3C Attribution & Publication Policy v 1.2-2020-08-25b supported by NCATS Contract No. 75N95023D00001, Axle Informatics Subcontract: NCATS-P00438-B, and [insert additional funding agencies or sources and reference numbers as declared by the contributors in their form response above]. This research was possible because of the patients whose information is included within the data and the organizations (https://ncats.nih.gov/n3c/resources/data-contribution/data-transfer-agreement-signatories) and scientists who have contributed to the on-going development of this community resource [https://doi.org/10.1093/jamia/ocaa196].

Disclaimer The N3C Publication committee confirmed that this white paper msid:1865.948 is in accordance with N3C data use and attribution policies; however, this content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the N3C program.

IRB

The N3C data transfer to NCATS is performed under a Johns Hopkins University Reliance Protocol # IRB00249128 or individual site agreements with NIH. The N3C Data Enclave is managed under the authority of the NIH; information can be found at https://ncats.nih.gov/n3c/resources.

Individual Acknowledgements for Core Contributors

We gratefully acknowledge the following core contributors to N3C:

Adam B. Wilcox, Adam M. Lee, Alexis Graves, Alfred (Jerrod) Anzalone, Amin Manna, Amit Saha, Amy Olex, Andrea Zhou, Andrew E. Williams, Andrew M. Southerland, Andrew T. Girvin, Anita Walden, Anjali Sharathkumar, Benjamin Amor, Benjamin Bates, Brian Hendricks, Brijesh Patel, G. Caleb Alexander, Carolyn T. Bramante, Cavin Ward-Caviness, Charisse Madlock-Brown, Christine Suver, Christopher G. Chute, Christopher Dillon, Chunlei Wu, Clare Schmitt, Cliff Takemoto, Dan Housman, Davera Gabriel, David A. Eichmann, Diego Mazzotti, Donald E. Brown, Eilis Boudreau, Elaine L. Hill, Emily Carlson Marti, Emily R. Pfaff, Evan French, Farrukh M Koraishy, Federico Mariona, Fred Prior, George Sokos, Greg Martin, Harold P. Lehmann, Heidi Spratt, Hemalkumar B. Mehta, J.W. Awori Hayanga, Jami Pincavitch, Jaylyn Clark, Jeremy Richard Harper, Jessica Yasmine Islam, Jin Ge, Joel Gagnier, Johanna J. Loomba, John B. Buse, Jomol Mathew, Joni L. Rutter, Julie A. McMurry, Justin Guinney, Justin Starren, Karen Crowley, Katie Rebecca Bradwell, Kellie M. Walters, Ken Wilkins, Kenneth R. Gersing, Kenrick Cato, Kimberly Murray, Kristin Kostka, Lavance Northington, Lee Pyles, Lesley Cottrell, Lili M. Portilla, Mariam Deacy, Mark M. Bissell, Marshall Clark, Mary Emmett, Matvey B. Palchuk, Melissa A. Haendel, Meredith Adams, Meredith Temple-O'Connor, Michael G. Kurilla, Michele Morris, Nasia Safdar, Nicole Garbarini, Noha Sharafeldin, Ofer Sadan, Patricia A. Francis, Penny Wung Burgoon, Philip R.O. Payne, Randeep Jawa, Rebecca Erwin-Cohen, Rena C. Patel, Richard A. Moffitt, Richard L. Zhu, Rishikesan Kamaleswaran, Robert Hurley, Robert T. Miller, Saiju Pyarajan, Sam G. Michael, Samuel Bozzette, Sandeep K. Mallipattu, Satyanarayana Vedula, Scott Chapman, Shawn T. O'Neil, Soko Setoguchi, Stephanie S. Hong, Steven G. Johnson, Tellen D. Bennett, Tiffany J.

Callahan, Umit Topaloglu, Valery Gordon, Vignesh Subbian, Warren A. Kibbe, Wenndy Hernandez, Will Beasley, Will Cooper, William Hillegass, Xiaohan Tanner Zhang. Details of contributions available at covid.cd2h.org/core-contributors

Data Partners with Released Data

The following institutions whose data is released or pending:

Available: Advocate Health Care Network — UL1TR002389: The Institute for Translational Medicine (ITM) • Aurora Health Care Inc — UL1TR002373: Wisconsin Network For Health Research • Boston University Medical Campus — UL1TR001430: Boston University Clinical and Translational Science Institute • Brown University — U54GM115677: Advance Clinical Translational Research (Advance-CTR) • Carilion Clinic — UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • Case Western Reserve University — UL1TR002548: The Clinical & Translational Science Collaborative of Cleveland (CTSC) • Charleston Area Medical Center — U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) • Children's Hospital Colorado — UL1TR002535: Colorado Clinical and Translational Sciences Institute • Columbia University Irving Medical Center — UL1TR001873: Irving Institute for Clinical and Translational Research • Dartmouth College — None (Voluntary) Duke University — UL1TR002553: Duke Clinical and Translational Science Institute • George Washington Children's Research Institute — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • George Washington University — UL1TR001876: Clinical and Translational Science Institute at Children's National (CTSA-CN) • Harvard Medical School — UL1TR002541: Harvard Catalyst • Indiana University School of Medicine — UL1TR002529: Indiana Clinical and Translational Science Institute • Johns Hopkins University — UL1TR003098: Johns Hopkins Institute for Clinical and Translational Research • Louisiana Public Health Institute — None (Voluntary) • Loyola Medicine — Loyola University Medical Center • Loyola University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • Maine Medical Center — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • Mary Hitchcock Memorial Hospital & Dartmouth Hitchcock Clinic — None (Voluntary) • Massachusetts General Brigham — UL1TR002541: Harvard Catalyst • Mayo Clinic Rochester — UL1TR002377: Mayo Clinic Center for Clinical and Translational Science (CCaTS) • Medical University of South Carolina — UL1TR001450: South Carolina Clinical & Translational Research Institute (SCTR) • MITRE Corporation — None (Voluntary) • Montefiore Medical Center — UL1TR002556: Institute for Clinical and Translational Research at Einstein and Montefiore • Nemours — U54GM104941: Delaware CTR ACCEL Program • NorthShore University HealthSystem — UL1TR002389: The Institute for Translational Medicine (ITM) • Northwestern University at Chicago — UL1TR001422: Northwestern University Clinical and Translational Science Institute (NUCATS) • OCHIN — INV-018455: Bill and Melinda Gates Foundation grant to Sage Bionetworks • Oregon Health & Science University — UL1TR002369: Oregon Clinical and Translational Research Institute • Penn State Health Milton S. Hershey Medical Center — UL1TR002014: Penn State Clinical and Translational Science Institute • Rush University Medical Center — UL1TR002389: The Institute for Translational Medicine (ITM) • Rutgers, The State University of New Jersey — UL1TR003017: New Jersey Alliance for Clinical and Translational Science • Stony Brook University — U24TR002306 • The Alliance at the University of Puerto Rico, Medical Sciences Campus — U54GM133807: Hispanic Alliance for Clinical and Translational Research (The Alliance) • The Ohio State University — UL1TR002733: Center for Clinical and Translational Science • The State University of New York at Buffalo — UL1TR001412: Clinical and Translational Science Institute • The University of Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • The University of Iowa — UL1TR002537: Institute for Clinical and Translational Science • The University of Miami Leonard M. Miller School of Medicine — UL1TR002736: University of Miami Clinical and Translational Science

Institute • The University of Michigan at Ann Arbor — UL1TR002240: Michigan Institute for Clinical and Health Research • The University of Texas Health Science Center at Houston — UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • The University of Texas Medical Branch at Galveston — UL1TR001439: The Institute for Translational Sciences • The University of Utah — UL1TR002538: Uhealth Center for Clinical and Translational Science • Tufts Medical Center — UL1TR002544: Tufts Clinical and Translational Science Institute • Tulane University — UL1TR003096: Center for Clinical and Translational Science • The Queens Medical Center — None (Voluntary) • University Medical Center New Orleans — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • University of Alabama at Birmingham — UL1TR003096: Center for Clinical and Translational Science • University of Arkansas for Medical Sciences — UL1TR003107: UAMS Translational Research Institute • University of Cincinnati — UL1TR001425: Center for Clinical and Translational Science and Training • University of Colorado Denver, Anschutz Medical Campus — UL1TR002535: Colorado Clinical and Translational Sciences Institute • University of Illinois at Chicago — UL1TR002003: UIC Center for Clinical and Translational Science • University of Kansas Medical Center — UL1TR002366: Frontiers: University of Kansas Clinical and Translational Science Institute • University of Kentucky — UL1TR001998: UK Center for Clinical and Translational Science • University of Massachusetts Medical School Worcester — UL1TR001453: The UMass Center for Clinical and Translational Science (UMCCTS) • University Medical Center of Southern Nevada — None (voluntary) • University of Minnesota — UL1TR002494: Clinical and Translational Science Institute • University of Mississippi Medical Center — U54GM115428: Mississippi Center for Clinical and Translational Research (CCTR) • University of Nebraska Medical Center — U54GM115458: Great Plains IDeA-Clinical & Translational Research • University of North Carolina at Chapel Hill — UL1TR002489: North Carolina Translational and Clinical Science Institute • University of Oklahoma Health Sciences Center — U54GM104938: Oklahoma Clinical and Translational Science Institute (OCTSI) • University of Pittsburgh — UL1TR001857: The Clinical and Translational Science Institute (CTSI) • University of Pennsylvania — UL1TR001878: Institute for Translational Medicine and Therapeutics • University of Rochester — UL1TR002001: UR Clinical & Translational Science Institute • University of Southern California — UL1TR001855: The Southern California Clinical and Translational Science Institute (SC CTSI) • University of Vermont — U54GM115516: Northern New England Clinical & Translational Research (NNE-CTR) Network • University of Virginia — UL1TR003015: iTHRIV Integrated Translational health Research Institute of Virginia • University of Washington — UL1TR002319: Institute of Translational Health Sciences • University of Wisconsin-Madison — UL1TR002373: UW Institute for Clinical and Translational Research • Vanderbilt University Medical Center — UL1TR002243: Vanderbilt Institute for Clinical and Translational Research • Virginia Commonwealth University — UL1TR002649: C. Kenneth and Dianne Wright Center for Clinical and Translational Research • Wake Forest University Health Sciences — UL1TR001420: Wake Forest Clinical and Translational Science Institute • Washington University in St. Louis — UL1TR002345: Institute of Clinical and Translational Sciences • Weill Medical College of Cornell University — UL1TR002384: Weill Cornell Medicine Clinical and Translational Science Center • West Virginia University — U54GM104942: West Virginia Clinical and Translational Science Institute (WVCTSI) Submitted: Icahn School of Medicine at Mount Sinai — UL1TR001433: ConduITS Institute for Translational Sciences • The University of Texas Health Science Center at Tyler — UL1TR003167: Center for Clinical and Translational Sciences (CCTS) • University of California, Davis — UL1TR001860: UCDavis Health Clinical and Translational Science Center • University of California, Irvine — UL1TR001414: The UC Irvine Institute for Clinical and Translational Science (ICTS) • University of California, Los Angeles — UL1TR001881: UCLA Clinical Translational Science Institute • University of California, San Diego — UL1TR001442: Altman Clinical and Translational Research Institute • University of California, San Francisco — UL1TR001872: UCSF Clinical and Translational Science Institute NYU Langone Health Clinical Science Core, Data Resource Core, and PASC Biorepository Core —

OTA-21-015A: Post-Acute Sequelae of SARS-CoV-2 Infection Initiative (RECOVER) Pending: Arkansas Children's Hospital — UL1TR003107: UAMS Translational Research Institute • Baylor College of Medicine — None (Voluntary) • Children's Hospital of Philadelphia — UL1TR001878: Institute for Translational Medicine and Therapeutics • Cincinnati Children's Hospital Medical Center — UL1TR001425: Center for Clinical and Translational Science and Training • Emory University — UL1TR002378: Georgia Clinical and Translational Science Alliance • HonorHealth — None (Voluntary) • Loyola University Chicago — UL1TR002389: The Institute for Translational Medicine (ITM) • Medical College of Wisconsin — UL1TR001436: Clinical and Translational Science Institute of Southeast Wisconsin • MedStar Health Research Institute — None (Voluntary) • Georgetown University — UL1TR001409: The Georgetown-Howard Universities Center for Clinical and Translational Science (GHUCCTS) • MetroHealth — None (Voluntary) • Montana State University — U54GM115371: American Indian/Alaska Native CTR • NYU Langone Medical Center — UL1TR001445: Langone Health's Clinical and Translational Science Institute • Ochsner Medical Center — U54GM104940: Louisiana Clinical and Translational Science (LA CaTS) Center • Regenstrief Institute — UL1TR002529: Indiana Clinical and Translational Science Institute • Sanford Research — None (Voluntary) • Stanford University — UL1TR003142: Spectrum: The Stanford Center for Clinical and Translational Research and Education • The Rockefeller University — UL1TR001866: Center for Clinical and Translational Science • The Scripps Research Institute — UL1TR002550: Scripps Research Translational Institute • University of Florida — UL1TR001427: UF Clinical and Translational Science Institute • University of New Mexico Health Sciences Center — UL1TR001449: University of New Mexico Clinical and Translational Science Center • University of Texas Health Science Center at San Antonio — UL1TR002645: Institute for Integration of Medicine and Science • Yale New Haven Hospital — UL1TR001863: Yale Center for Clinical Investigation

Tables

Table 1. Fairness metrics for baseline and after bias mitigation (using reweighting, MAAT, and FairMask), optimizing for a single protected attribute (sex, race, or ethnicity)

Model type	Protected attribute	Disparate impact ratio (DIR)				Equal opportunity ratio (EOR)				Predictive equality ratio (PER)			
		Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
Logistic Regression	Sex	0.17	0.91 ↑	0.17	0.89 ↑	0.31	0.98 ↑	0.31	0.96 ↑	0.16	0.92 ↑	0.16	0.89 ↑
	Race	0.86	0.84 ↓	0.85 ↓	0.79 ↓	0.92	0.93 ↑	0.91 ↓	0.80 ↓	0.85	0.84 ↓	0.85	0.79 ↓
	Ethnicity	0.90	0.91 ↑	0.90	0.93 ↑	0.93	0.94 ↑	0.93	0.95 ↑	0.89	0.91 ↑	0.89	0.93 ↑
Random Forest	Sex	0.56	0.93 ↑	0.56	0.92 ↑	0.71	0.97 ↑	0.71	0.98 ↑	0.56	0.93 ↑	0.56	0.92 ↑
	Race	0.87	0.88 ↑	0.87	0.76 ↓	0.85	0.86 ↑	0.85	0.79 ↓	0.87	0.88 ↑	0.87	0.76 ↓
	Ethnicity	0.88	0.89 ↑	0.88	0.84 ↓	0.91	0.91	0.91	0.88 ↓	0.88	0.89 ↑	0.88	0.84 ↓

Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specification. For bias mitigation optimizing for a single protected attribute, we specified the unprivileged group as the second largest non-missing category of the protected attribute: sex (male), race (Black or African American), and ethnicity (Hispanic or Latino). The Baseline columns represent the model fairness before any bias mitigation techniques are applied.

↑ increase from baseline

↓ decrease from baseline

MAAT = mitigating algorithmic bias with adversarial training.

Table 2. Performance metrics by patient subgroup at baseline and after bias mitigation (using reweighting, MAAT, and FairMask), optimizing for a single protected attribute (sex, race, or ethnicity)

			AUROC				PRAUC			
Model type	**Protected attribute**	**Demographic subgroup**	**Baseline**	**Reweighting**	**MAAT**	**FairMask**	**Baseline**	**Reweighting**	**MAAT**	**FairMask**
Logistic Regression	Sex	All	0.60	0.59 ↓	0.60	0.59 ↓	0.35	0.28 ↓	0.35	0.31 ↓
		Female	0.57	0.58 ↑	0.57	0.58 ↑	0.43	0.29 ↓	0.43	0.32 ↓
		Male	0.57	0.60 ↑	0.57	0.59 ↑	0.17	0.27 ↑	0.17	0.29 ↑
		Other	0.56	0.64 ↑	0.56	0.60 ↑	0.26	0.25 ↓	0.27 ↑	0.25 ↓
	Race	All	0.60	0.60	0.60	0.60	0.35	0.35	0.35	0.36 ↑
		White	0.59	0.59	0.59	0.59	0.37	0.37	0.37	0.37
		Black or African American	0.62	0.62	0.62	0.62	0.35	0.34 ↓	0.35	0.42 ↑
		Asian	0.57	0.58 ↑	0.58 ↑	0.58 ↑	0.21	0.21	0.21	0.21
		NHPI	0.56	0.56	0.57 ↑	0.57 ↑	0.19	0.19	0.20 ↑	0.20 ↑
		Other	0.62	0.62	0.63 ↑	0.62	0.34	0.33 ↓	0.34	0.33 ↓
		Missing or unknown	0.60	0.60	0.60	0.60	0.23	0.23	0.23	0.25 ↑
	Ethnicity	All	0.60	0.60	0.60	0.60	0.35	0.35	0.35	0.36 ↑
		Not Hispanic or Latino	0.60	0.60	0.60	0.60	0.37	0.37	0.37	0.37
		Hispanic or Latino	0.60	0.60	0.60	0.60	0.33	0.33	0.33	0.35 ↑
		Missing or unknown	0.59	0.59	0.59	0.61 ↑	0.20	0.20	0.20	0.24 ↑
Random Forest	Sex	All	0.60	0.60	0.60	0.60	0.33	0.30 ↓	0.33	0.30 ↓
		Female	0.59	0.59	0.59	0.59	0.36	0.31 ↓	0.36	0.31 ↓
		Male	0.60	0.61 ↑	0.60	0.61 ↑	0.26	0.30 ↑	0.26	0.29 ↑
		Other	0.89	0.81 ↓	0.89	0.83 ↓	0.54	0.42 ↓	0.54	0.45 ↓
	Race	All	0.60	0.60	0.60	0.60	0.33	0.33	0.33	0.34 ↑
		White	0.59	0.59	0.59	0.60 ↑	0.33	0.33	0.33	0.34 ↑
		Black or African American	0.62	0.62	0.62	0.62	0.37	0.37	0.37	0.39 ↑
		Asian	0.57	0.58 ↑	0.57	0.56 ↓	0.25	0.25	0.25	0.22 ↓
		NHPI	0.55	0.55	0.55	0.55	0.22	0.22	0.22	0.21 ↓
		Other	0.60	0.60	0.60	0.58 ↓	0.29	0.29	0.29	0.24 ↓
		Missing or Unknown	0.62	0.62	0.62	0.61 ↓	0.27	0.27	0.27	0.26 ↓
	Ethnicity	All	0.60	0.60	0.60	0.60	0.33	0.33	0.33	0.33
		Not Hispanic or Latino	0.60	0.60	0.60	0.60	0.34	0.34	0.34	0.34
		Hispanic or Latino	0.60	0.60	0.60	0.60	0.31	0.31	0.31	0.30 ↓
		Missing or Unknown	0.62	0.62	0.62	0.62	0.25	0.25	0.25	0.25

Table 2 (*continued*)

Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specification. For bias mitigation optimizing for a single protected attribute, we specified the unprivileged group as the second largest non-missing category of the protected attribute: sex (male), race (Black or African American), and ethnicity (Hispanic or Latino). The "All" rows are the AUROC and PRAUC for the model across demographic subgroups. The Baseline columns represent model performance before any bias mitigation techniques are applied.

↑ increase from baseline

↓ decrease from baseline

AUROC = area under the receiver operating characteristic curve; MAAT = mitigating algorithmic bias with adversarial training; NHPI = Native Hawaiian or Pacific Islander; PRAUC = area under the precision-recall curve.

Table 3. Fairness metrics for fatigue symptom cluster before and after bias mitigation when optimizing multiple attributes

Model type	Protected characteristic	Disparate impact ratio				Equal opportunity ratio				Predictive equality ratio			
		Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
LR	Sex	0.17	0.25 ↑	0.17	0.90 ↑	0.31	0.41 ↑	0.31	0.99 ↑	0.16	0.24 ↑	0.16	0.90 ↑
	Race	0.86	0.88 ↑	0.86	0.72 ↓	0.92	0.87 ↓	0.91 ↓	0.74 ↓	0.85	0.88 ↑	0.85	0.72 ↓
	Ethnicity	0.90	0.90	0.90	0.92 ↑	0.93	0.93	0.93	0.92 ↓	0.89	0.90 ↑	0.89	0.92 ↑
	Intersectional metric	0.63	0.65 ↑	0.63	0.18 ↓	0.74	0.70 ↓	0.74	0.26 ↓	0.63	0.65 ↑	0.63	0.17 ↓
RF	Sex	0.56	0.50 ↓	0.56	0.92 ↑	0.71	0.68 ↓	0.71	0.98 ↑	0.56	0.50 ↓	0.56	0.92 ↑
	Race	0.87	0.82 ↓	0.87	0.80 ↓	0.85	0.84 ↓	0.85	0.78 ↓	0.87	0.81 ↓	0.87	0.80 ↓
	Ethnicity	0.88	0.85 ↓	0.88	0.85 ↓	0.91	0.89 ↓	0.91	0.89 ↓	0.88	0.85 ↓	0.88	0.85 ↓
	Intersectional metric	0.35	0.55 ↑	0.35	0.18 ↓	0.40	0.54 ↑	0.40	0.27 ↓	0.34	0.55 ↑	0.34	0.17 ↓

Source Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specifications after optimizing the combination of all protected attributes. We specified the protected attribute as the second largest non-missing category for the demographic categories sex (male), race (Black or African American), and ethnicity (Hispanic or Latino). The Baseline columns represent model fairness before any bias mitigation techniques are applied.

↑ increase from baseline

↓ decrease from baseline

LR = Logistic Regression; MAAT = mitigating algorithmic bias with adversarial training; RF = Random Forest.

Table 4. Performance metrics for fatigue symptom cluster before and after bias mitigation when optimizing multiple attributes

Model type	Demographic subgroup	AUROC				PRAUC			
		Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
LR	All	0.60	0.60	0.60	0.59 ↓	0.35	0.38 ↑	0.35	0.29 ↓
	Sex: Female	0.57	0.56 ↓	0.57	0.58 ↑	0.43	0.45 ↑	0.43	0.30 ↓
	Sex: Male	0.57	0.58 ↑	0.57	0.60 ↑	0.17	0.21 ↑	0.17	0.27 ↑
	Sex: Other	0.56	0.55 ↓	0.56	0.61 ↑	0.26	0.27 ↑	0.27 ↑	0.25 ↓
	Race: Asian	0.57	0.58 ↑	0.58 ↑	0.55 ↓	0.21	0.22 ↑	0.21	0.16 ↓
	Race: Black or African American	0.62	0.62	0.62	0.61 ↓	0.35	0.42 ↑	0.35	0.35
	Race: Missing or unknown	0.60	0.61 ↑	0.60	0.59 ↓	0.23	0.29 ↑	0.23	0.22 ↓
	Race: NHPI	0.56	0.56	0.57 ↑	0.55 ↓	0.19	0.20 ↑	0.20 ↑	0.17 ↓
	Race: Other	0.62	0.63 ↑	0.63 ↑	0.58 ↓	0.34	0.36 ↑	0.34	0.30 ↓
	Race: White	0.59	0.59	0.59	0.58 ↓	0.37	0.39 ↑	0.37	0.29 ↓
	Ethnicity: Hispanic or Latino	0.60	0.60	0.60	0.58 ↓	0.33	0.40 ↑	0.33	0.27 ↓
	Ethnicity: Missing or unknown	0.59	0.60 ↑	0.59	0.58 ↓	0.20	0.21 ↑	0.20	0.19 ↓
	Ethnicity: Not Hispanic or Latino	0.60	0.60	0.60	0.59 ↓	0.37	0.39 ↑	0.37	0.30 ↓
RF	All	0.60	0.60	0.60	0.60	0.33	0.37 ↑	0.33	0.30 ↓
	Sex: Female	0.59	0.57 ↓	0.59	0.59	0.36	0.41 ↑	0.36	0.31 ↓
	Sex: Male	0.60	0.61 ↑	0.60	0.61 ↑	0.26	0.29 ↑	0.26	0.30 ↑
	Sex: Other	0.89	0.85 ↓	0.89	0.75 ↓	0.54	0.53 ↓	0.54	0.36 ↓
	Race: Asian	0.57	0.59 ↑	0.57	0.57	0.25	0.29 ↑	0.25	0.24 ↓
	Race: Black or African American	0.62	0.61 ↓	0.62	0.62	0.37	0.42 ↑	0.37	0.36 ↓
	Race: Missing or unknown	0.62	0.63 ↑	0.62	0.61 ↓	0.27	0.33 ↑	0.27	0.26 ↓
	Race: NHPI	0.55	0.56 ↑	0.55	0.54 ↓	0.22	0.27 ↑	0.22	0.21 ↓
	Race: Other	0.60	0.64 ↑	0.60	0.59 ↓	0.29	0.36 ↑	0.29	0.25 ↓
	Race: White	0.59	0.59	0.59	0.59	0.33	0.37 ↑	0.33	0.30 ↓
	Ethnicity: Hispanic or Latino	0.60	0.60	0.60	0.59 ↓	0.31	0.40 ↑	0.31	0.28 ↓
	Ethnicity: Missing or unknown	0.62	0.63 ↑	0.62	0.62	0.25	0.28 ↑	0.25	0.26 ↑
	Ethnicity: Not Hispanic or Latino	0.60	0.60	0.60	0.59 ↓	0.34	0.37 ↑	0.34	0.31 ↓

Table 4 (*continued*)

Source Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Notes: Results are shown as the average across 100 bootstrap samples for each model specifications when optimizing all attributes. We specified the protected attribute as the second largest non-missing category for the demographic categories sex (male), race (Black or African American), and ethnicity (Hispanic or Latino). The demographic subgroups are shown in accordance with the demographic characteristics on which bias mitigation was performed. "All" is the AUROC and PRAUC for the model across characteristic subgroups. The Baseline columns represent model performance before any bias mitigation techniques are applied.

↑ increase from baseline

↓ decrease from baseline

AUROC: area under the receiver operating characteristic curve; LR = Logistic Regression; MAAT = mitigating algorithmic bias with adversarial training; NHPI = Native Hawaiian or Pacific Islander; PRAUC = area under the precision-recall curve; RF = Random Forest.

Appendix

Table A.1. Analytic sample

Demographic characteristics	Cognitive symptom cluster		Fatigue symptom cluster		Respiratory symptom cluster	
	Positive N (%)	Negative N (%)	Positive N (%)	Negative N (%)	Positive N (%)	Negative N (%)
Total	2,937 (0.2%)	1,231,182 (99.8%)	57,483 (4.7%)	1,176,636 (95.3%)	30,668 (2.5%)	1,203,451 (97.5%)
Age	60.1 (SD: 17.22)	41.4 (SD: 20.72)	44.1 (SD: 19.2)	41.3 (SD: 20.8)	47.0 (SD: 22)	41.3 (SD: 20.68)
Age group						
0-17	40 (<0.1%)	171,327 (>99.9%)	4,597 (2.7%)	166,770 (97.3%)	3,683 (2.1%)	167,684 (97.9%)
18-29	126 (<0.1%)	218,308 (>99.9%)	10,365 (4.7%)	208,069 (95.3%)	3,327 (1.5%)	215,107 (98.5%)
30-39	179 (0.1%)	188,841 (99.9%)	9,583 (5.1%)	179,437 (94.9%)	3,409 (1.8%)	185,611 (98.2%)
40-49	451 (0.2%)	194,235 (99.8%)	10,029 (5.2%)	184,657 (94.8%)	4,561 (2.3%)	190,125 (97.7%)
50-59	474 (0.3%)	186,883 (99.7%)	9,173 (4.9%)	178,184 (95.1%)	5,407 (2.9%)	181,950 (97.1%)
60-69	623 (0.4%)	150,740 (99.6%)	7,393 (4.9%)	143,970 (95.1%)	5,297 (3.5%)	146,066 (96.5%)
70-79	681 (0.8%)	87,738 (99.2%)	4,597 (5.2%)	83,822 (94.8%)	3,665 (4.1%)	84,754 (95.9%)
80+	363 (1.1%)	33,110 (98.9%)	1,746 (5.2%)	31,727 (94.8%)	1,319 (3.9%)	32,154 (96.1%)
Sex						
Female	1,751 (0.3%)	690,796 (99.7%)	39,362 (5.7%)	653,185 (94.3%)	18,164 (2.6%)	674,383 (97.4%)
Male	1,186 (0.2%)	540,035 (99.8%)	18,111 ‡ (3.3%)	523,112 (96.7%)	12,501‡ (2.3%)	528,718 (97.7%)
Other	0 (0%)	351 (100%)	< 20 †	337 ‡	< 20 †	352 ‡
Race						
Asian	56 (0.2%)	25,719 (99.8%)	929 (3.6%)	24,846 (96.4%)	619 (2.4%)	25,156 (97.6%)
Black or African American	412 (0.2%)	172,825 (99.8%)	8,068 (4.7%)	165,169 (95.3%)	5,186 (3%)	168,051 (97%)
NHPI	< 20 †	1,866 ‡	61 (3.3%)	1,805 (96.7%)	38 (2%)	1,828 (98%)
Other	< 20 †	10,393 ‡	470 (4.5%)	9,937 (95.5%)	286 (2.7%)	10,121 (97.3%)
White	2,163 (0.3%)	845,108 (99.7%)	41,449 (4.9%)	805,822 (95.1%)	21,293 (2.5%)	825,978 (97.5%)
Missing or unknown	289 (0.2%)	175,274 (99.8%)	6,506 (3.7%)	169,057 (96.3%)	3,246 (1.8%)	172,317 (98.2%)
Ethnicity						
Hispanic or Latino	295 (0.2%)	167,080 (99.8%)	7,691 (4.6%)	159,684 (95.4%)	3,664 (2.2%)	163,711 (97.8%)
Not Hispanic or Latino	2,408 (0.3%)	946,774 (99.7%)	45,876 (4.8%)	903,306 (95.2%)	24,966 (2.6%)	924,216 (97.4%)
Missing or unknown	234 (0.2%)	117,328 (99.8%)	3,916 (3.3%)	113,646 (96.7%)	2,038 (1.7%)	115,524 (98.3%)
Smoking status						
Current or former smoker	529 (0.4%)	146,835 (99.6%)	9,472 (6.4%)	137,892 (93.6%)	5,486 (3.7%)	141,878 (96.3%)
Non-smoker	2,408 (0.2%)	1,084,347 (99.8%)	48,011 (4.4%)	1,038,744 (95.6%)	25,182 (2.3%)	1,061,573 (97.7%)
Hypertension	29 (0.7%)	4,101 (99.3%)	329 (8.0%)	3,801 (92.0%)	172 (4.2%)	3,958 (95.8%)

Table A.1. (*continued*)

Demographic characteristics	Cognitive symptom cluster		Fatigue symptom cluster		Respiratory symptom cluster	
	Positive N (%)	Negative N (%)	Positive N (%)	Negative N (%)	Positive N (%)	Negative N (%)
Obesity	590 (0.5%)	115,173 (99.5%)	9,581 (8.3%)	106,182 (91.7%)	5,281 (4.6%)	110,482 (95.4%)
CCI Score	2.7 (SD: 3.11)	0.8 (SD: 1.73)	1.4 (SD: 2.35)	0.7 (SD: 1.7)	1.6 (SD: 2.51)	0.7 (SD: 1.71)

Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Analytic sample of patients in N3C de-identified "tier 2 access" data set with lab confirmed positive cases of COVID-19 between September 1, 2020, and September 1, 2021, with condition occurrences within the six months preceding and proceeding diagnosis of COVID-19. The long symptom clusters cognitive, fatigue, and respiratory are defined by the Global Burden of Disease Long COVID Collaborators.

† To comply with N3C policy, counts below 20 are displayed as < 20, and in this case, additional values must be skewed by up to five to render it impossible to back-calculate precise counts fewer than 20 for the following categories: Age Group 0-9, Sex Other, Native Hawaiian or Pacific Islander, Race Other, and Pregnant.

‡ This proportion is one of the two columns that sum up to one. Reporting it would enable the calculation of a cell size < 20. Therefore, we mark it as too small to quantitatively report.

CCI =Charlson Comorbidity Index; NHPI = Native Hawaiian or Pacific Islander; N3C = National COVID Cohort Collaborative; SD = standard deviation.

Table A.2. Fairness metrics at baseline and after bias mitigation (using reweighting, MAAT, and FairMask), optimizing for a single protected variable (sex, race, or ethnicity) for the three symptom clusters.

Symptom cluster	Model type	Protected attribute	Disparate impact ratio (DIR)				Equal opportunity ratio (EOR)				Predictive equality ratio (PER)			
			Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
Cognitive	Logistic Regression	Sex	0.79	0.90 ↑	0.78 ↓	0.94 ↑	0.97	0.97	0.97	0.92 ↓	0.79	0.90 ↑	0.78 ↓	0.94 ↑
		Race	0.90	0.92 ↑	0.90	0.95 ↑	0.90	0.89 ↓	0.90	0.92 ↑	0.90	0.92 ↑	0.90	0.95 ↑
		Ethnicity	0.57	0.88 ↑	0.57	0.61 ↑	0.86	0.97 ↑	0.87 ↑	0.88 ↑	0.57	0.88 ↑	0.57	0.61 ↑
	Random Forest	Sex	0.98	0.99 ↑	0.98	0.99 ↑	0.98	0.97 ↓	0.98	0.97 ↓	0.98	0.99 ↑	0.98	0.99 ↑
		Race	0.90	0.91 ↑	0.90	0.91 ↑	0.91	0.91	0.91	0.91	0.90	0.90	0.90	0.90
		Ethnicity	0.70	0.74 ↑	0.70	0.70	0.91	0.95 ↑	0.91	0.90 ↓	0.70	0.74 ↑	0.70	0.69 ↓
Fatigue	Logistic Regression	Sex	0.17	0.91 ↑	0.17	0.89 ↑	0.31	0.98 ↑	0.31	0.96 ↑	0.16	0.92 ↑	0.16	0.89 ↑
		Race	0.86	0.84 ↓	0.85 ↓	0.79 ↓	0.92	0.93 ↑	0.91 ↓	0.80 ↓	0.85	0.84 ↓	0.85	0.79 ↓
		Ethnicity	0.90	0.91 ↑	0.90	0.93 ↑	0.93	0.94 ↑	0.93	0.95 ↑	0.89	0.91 ↑	0.89	0.93 ↑
	Random Forest	Sex	0.56	0.93 ↑	0.56	0.92 ↑	0.71	0.97 ↑	0.71	0.98 ↑	0.56	0.93 ↑	0.56	0.92 ↑
		Race	0.87	0.88 ↑	0.87	0.76 ↓	0.85	0.86 ↑	0.85	0.79 ↓	0.87	0.88 ↑	0.87	0.76 ↓
		Ethnicity	0.88	0.89 ↑	0.88	0.84 ↓	0.91	0.91	0.91	0.88 ↓	0.88	0.89 ↑	0.88	0.84 ↓
Respiratory	Logistic Regression	Sex	0.68	0.92 ↑	0.68	0.97 ↑	0.82	0.97 ↑	0.82	0.95 ↑	0.68	0.92 ↑	0.68	0.97 ↑
		Race	0.64	0.94 ↑	0.64	0.83 ↑	0.76	0.96 ↑	0.76	0.89 ↑	0.64	0.94 ↑	0.64	0.83 ↑
		Ethnicity	0.66	0.94 ↑	0.66	0.74 ↑	0.72	0.95 ↑	0.72	0.77 ↑	0.66	0.94 ↑	0.66	0.74 ↑
	Random Forest	Sex	0.96	0.98 ↑	0.96	0.98 ↑	0.95	0.97 ↑	0.95	0.97 ↑	0.96	0.98 ↑	0.96	0.98 ↑
		Race	0.76	0.82 ↑	0.76	0.79 ↑	0.85	0.88 ↑	0.85	0.87 ↑	0.76	0.82 ↑	0.76	0.79 ↑
		Ethnicity	0.80	0.87 ↑	0.80	0.84 ↑	0.79	0.84 ↑	0.79	0.82 ↑	0.81	0.87 ↑	0.81	0.85 ↑

Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specifications. We specified the protected attribute as the second largest non-missing category for the demographic categories sex (male), race (Black or African American), and ethnicity (Hispanic or Latino). The Baseline columns represent model fairness before any bias mitigation techniques are applied.

↑ increase from baseline

↓ decrease from baseline

MAAT = mitigating algorithmic bias with adversarial training.

Table A.3. Performance metrics by demographic subgroup at baseline and after bias mitigation (using reweighting, MAAT, and FairMask), optimizing for a single protected variable (sex, race, or ethnicity) for the three symptom clusters

	Model	Protected		AUROC				PRAUC			
Symptom	type	attribute	Demographic subgroup	Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
Cognitive	LR	Sex	All	0.73	0.73	0.72 ↓	0.73	0.36	0.36	0.36	0.36
			Female	0.71	0.72 ↑	0.71	0.72 ↑	0.36	0.36	0.36	0.35 ↓
			Male	0.74	0.74	0.74	0.74	0.36	0.36	0.36	0.38 ↑
			Other	--	--	--	--	--	--	--	--
		Race	All	0.73	0.73	0.72 ↓	0.73	0.36	0.36	0.36	0.36
			White	0.73	0.73	0.73	0.73	0.37	0.37	0.38 ↑	0.37
			Black or African American	0.70	0.70	0.70	0.70	0.33	0.33	0.33	0.33
			Asian	0.72	0.72	0.72	0.73 ↑	0.36	0.35 ↓	0.35 ↓	0.36
			NHPI	--	--	--	--	--	--	--	--
			Other	0.83	0.83	0.83	0.82 ↓	0.42	0.42	0.42	0.41 ↓
			Missing or unknown	0.71	0.71	0.71	0.71	0.29	0.29	0.29	0.29
		Ethnicity	All	0.73	0.72 ↓	0.72 ↓	0.73	0.36	0.36	0.36	0.36
			Not Hispanic or Latino	0.72	0.72	0.72	0.72	0.37	0.36 ↓	0.37	0.37
			Hispanic or Latino	0.73	0.74 ↑	0.74 ↑	0.73	0.32	0.36 ↑	0.32	0.32
			Missing or unknown	0.73	0.73	0.73	0.73	0.32	0.31 ↓	0.32	0.32
	RF	Sex	All	0.72	0.72	0.72	0.72	0.37	0.37	0.37	0.37
			Female	0.72	0.72	0.72	0.72	0.37	0.37	0.37	0.37
			Male	0.73	0.73	0.73	0.73	0.38	0.38	0.38	0.38
			Other	--	--	--	--	--	--	--	--
		Race	All	0.72	0.72	0.72	0.72	0.37	0.37	0.37	0.37
			White	0.72	0.72	0.72	0.72	0.38	0.38	0.38	0.38
			Black or African American	0.68	0.68	0.68	0.68	0.34	0.34	0.34	0.34
			Asian	0.74	0.74	0.74	0.74	0.38	0.38	0.38	0.38
			NHPI	--	--	--	--	--	--	--	--
			Other	0.83	0.83	0.83	0.83	0.42	0.42	0.42	0.42
			Missing or unknown	0.74	0.74	0.74	0.73 ↓	0.34	0.34	0.34	0.34
		Ethnicity	All	0.72	0.72	0.72	0.72	0.37	0.37	0.37	0.37
			Not Hispanic or Latino	0.72	0.72	0.72	0.72	0.38	0.38	0.38	0.38
			Hispanic or Latino	0.73	0.74 ↑	0.73	0.73	0.34	0.35 ↑	0.34	0.34
			Missing or unknown	0.76	0.76	0.76	0.76	0.36	0.36	0.36	0.36

Table A.3. (*continued*)

Symptom	Model type	Protected attribute	Demographic subgroup	AUROC				PRAUC			
				Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
Fatigue	LR	Sex	All	0.60	0.59 ↓	0.60	0.59 ↓	0.35	0.28 ↓	0.35	0.31 ↓
			Female	0.57	0.58 ↑	0.57	0.58 ↑	0.43	0.29 ↓	0.43	0.32 ↓
			Male	0.57	0.60 ↑	0.57	0.59 ↑	0.17	0.27 ↑	0.17	0.29 ↑
			Other	0.56	0.64 ↑	0.56	0.60 ↑	0.26	0.25 ↓	0.27 ↑	0.25 ↓
		Race	All	0.60	0.60	0.60	0.60	0.35	0.35	0.35	0.36 ↑
			White	0.59	0.59	0.59	0.59	0.37	0.37	0.37	0.37
			Black or African American	0.62	0.62	0.62	0.62	0.35	0.34 ↓	0.35	0.42 ↑
			Asian	0.57	0.58 ↑	0.58 ↑	0.58 ↑	0.21	0.21	0.21	0.21
			NHPI	0.56	0.56	0.57 ↑	0.57 ↑	0.19	0.19	0.20 ↑	0.20 ↑
			Other	0.62	0.62	0.63 ↑	0.62	0.34	0.33 ↓	0.34	0.33 ↓
			Missing or unknown	0.60	0.60	0.60	0.60	0.23	0.23	0.23	0.25 ↑
		Ethnicity	All	0.60	0.60	0.60	0.60	0.35	0.35	0.35	0.36 ↑
			Not Hispanic or Latino	0.60	0.60	0.60	0.60	0.37	0.37	0.37	0.37
			Hispanic or Latino	0.60	0.60	0.60	0.60	0.33	0.33	0.33	0.35 ↑
			Missing or unknown	0.59	0.59	0.59	0.61 ↑	0.20	0.20	0.20	0.24 ↑
	RF	Sex	All	0.60	0.60	0.60	0.60	0.33	0.30 ↓	0.33	0.30 ↓
			Female	0.59	0.59	0.59	0.59	0.36	0.31 ↓	0.36	0.31 ↓
			Male	0.60	0.61 ↑	0.60	0.61 ↑	0.26	0.30 ↑	0.26	0.29 ↑
			Other	0.89	0.81 ↓	0.89	0.83 ↓	0.54	0.42 ↓	0.54	0.45 ↓
		Race	All	0.60	0.60	0.60	0.60	0.33	0.33	0.33	0.34 ↑
			White	0.59	0.59	0.59	0.60 ↑	0.33	0.33	0.33	0.34 ↑
			Black or African American	0.62	0.62	0.62	0.62	0.37	0.37	0.37	0.39 ↑
			Asian	0.57	0.58 ↑	0.57	0.56 ↓	0.25	0.25	0.25	0.22 ↓
			NHPI	0.55	0.55	0.55	0.55	0.22	0.22	0.22	0.21 ↓
			Other	0.60	0.60	0.60	0.58 ↓	0.29	0.29	0.29	0.24 ↓
			Missing or unknown	0.62	0.62	0.62	0.61 ↓	0.27	0.27	0.27	0.26 ↓
		Ethnicity	All	0.60	0.60	0.60	0.60	0.33	0.33	0.33	0.33
			Not Hispanic or Latino	0.60	0.60	0.60	0.60	0.34	0.34	0.34	0.34
			Hispanic or Latino	0.60	0.60	0.60	0.60	0.31	0.31	0.31	0.30 ↓
			Missing or unknown	0.62	0.62	0.62	0.62	0.25	0.25	0.25	0.25

Table A.3. *(continued)*

Symptom	Model type	Protected attribute	Demographic subgroup	AUROC				PRAUC			
				Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
Respiratory	LR	Sex	All	0.61	0.61	0.61	0.61	0.29	0.28 ↓	0.29	0.28 ↓
			Female	0.61	0.61	0.60 ↓	0.61	0.31	0.29 ↓	0.31	0.28 ↓
			Male	0.61	0.61	0.61	0.61	0.26	0.28 ↑	0.26	0.29 ↑
			Other	--	--	--	--	--	--	--	--
		Race	All	0.61	0.61	0.61	0.61	0.29	0.28 ↓	0.29	0.29
			White	0.60	0.60	0.60	0.61 ↑	0.29	0.30 ↑	0.29	0.29
			Black or African American	0.60	0.61 ↑	0.60	0.61 ↑	0.35	0.28 ↓	0.36 ↑	0.31 ↓
			Asian	0.60	0.60	0.60	0.60	0.26	0.27 ↑	0.26	0.26
			NHPI	0.62	0.62	0.63 ↑	0.62	0.23	0.24 ↑	0.25 ↑	0.24 ↑
			Other	0.58	0.58	0.58	0.59 ↑	0.31	0.32 ↑	0.31	0.30 ↓
			Missing or unknown	0.60	0.61 ↑	0.60	0.60	0.20	0.21 ↑	0.20	0.20
		Ethnicity	All	0.61	0.61	0.61	0.61	0.29	0.29	0.29	0.29
			Not Hispanic or Latino	0.61	0.61	0.61	0.61	0.31	0.30 ↓	0.31	0.31
			Hispanic or Latino	0.59	0.60 ↑	0.59	0.59	0.22	0.27 ↑	0.22	0.23 ↑
			Missing or unknown	0.60	0.59 ↓	0.60	0.60	0.18	0.17 ↓	0.18	0.18
	RF	Sex	All	0.62	0.62	0.62	0.62	0.30	0.30	0.30	0.30
			Female	0.62	0.62	0.62	0.62	0.31	0.31	0.31	0.31
			Male	0.61	0.61	0.61	0.61	0.29	0.29	0.29	0.30 ↑
			Other	--	--	--	--	--	--	--	--
		Race	All	0.62	0.62	0.62	0.62	0.30	0.30	0.30	0.30
			White	0.62	0.61 ↓	0.62	0.62	0.30	0.30	0.30	0.30
			Black or African American	0.62	0.62	0.62	0.62	0.34	0.33 ↓	0.34	0.34
			Asian	0.61	0.61	0.61	0.61	0.28	0.28	0.28	0.28
			NHPI	0.61	0.61	0.61	0.61	0.25	0.25	0.25	0.25
			Other	0.57	0.57	0.57	0.57	0.26	0.26	0.26	0.25 ↓
			Missing or unknown	0.62	0.62	0.62	0.62	0.26	0.26	0.26	0.25 ↓
		Ethnicity	All	0.62	0.62	0.62	0.62	0.30	0.30	0.30	0.30
			Not Hispanic or Latino	0.62	0.62	0.62	0.62	0.31	0.31	0.31	0.31
			Hispanic or Latino	0.59	0.60 ↑	0.59	0.59	0.25	0.26 ↑	0.25	0.25
			Missing or unknown	0.62	0.62	0.62	0.62	0.25	0.25	0.25	0.25

Source: Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Note: Results are shown as the average across 100 bootstrap samples for each model specifications. We specified the protected attribute as the second largest non-missing category for the demographic categories sex (male), race (Black or African American), and ethnicity (Hispanic or Latino). The demographic subgroups are shown in accordance with the demographic characteristics on which bias mitigation was performed. The "All" rows are the AUROC and PRAUC for the model across demographic subgroups. The Baseline columns represent model performance before any bias mitigation techniques are applied.

Table A.3. (*continued*)

↑ increase from baseline

↓ decrease from baseline

-- unable to calculate the AUROC or PRAUC for some demographic characteristic subgroups due to small sample size.

AUROC = area under the receiver operating characteristic curve; LR = logistic regression; MAAT = mitigating algorithmic bias with adversarial training; NHPI = Native Hawaiian or Pacific Islander; PRAUC = area under the precision-recall curve; RF = Random Forest.

Table A.4. Mean model fairness for test data before and after bias mitigation when optimizing multiple attributes

Symptom cluster	Model type	Protected attribute	Disparate impact ratio				Equal opportunity ratio				Predictive equality ratio			
			Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
Cognitive	LR	Sex	0.79	0.83 ↑	0.78 ↓	0.95 ↑	0.97	0.98 ↑	0.97	0.93 ↓	0.79	0.83 ↑	0.78 ↓	0.95 ↑
		Race	0.90	0.90	0.90	0.96 ↑	0.90	0.91 ↑	0.90	0.91 ↑	0.90	0.90	0.90	0.96 ↑
		Ethnicity	0.57	0.60 ↑	0.57	0.63 ↑	0.86	0.88 ↑	0.87 ↑	0.92 ↑	0.57	0.60 ↑	0.57	0.63 ↑
		Intersectional metric	0.21	0.20 ↓	0.22 ↑	0.14 ↓	0.35	0.32 ↓	0.34 ↓	0.30 ↓	0.21	0.20 ↓	0.21	0.13
	RF	Sex	0.98	0.98	0.98	0.99 ↑	0.98	0.98	0.98	0.97 ↓	0.98	0.98	0.98	0.99 ↑
		Race	0.90	0.90	0.90	0.91 ↑	0.91	0.91	0.91	0.91	0.90	0.90	0.90	0.91 ↑
		Ethnicity	0.70	0.71 ↑	0.70	0.71 ↑	0.91	0.93 ↑	0.91	0.92 ↑	0.70	0.71 ↑	0.70	0.71 ↑
		Intersectional metric	0.16	0.16	0.16	0.14 ↓	0.34	0.34	0.34	0.34	0.16	0.16	0.16	0.14
Fatigue	LR	Sex	0.17	0.25 ↑	0.17	0.90 ↑	0.31	0.41 ↑	0.31	0.99 ↑	0.16	0.24 ↑	0.16	0.90 ↑
		Race	0.86	0.88 ↑	0.86	0.72 ↓	0.92	0.87 ↓	0.91 ↓	0.74 ↓	0.85	0.88 ↑	0.85	0.72 ↓
		Ethnicity	0.90	0.90	0.90	0.92 ↑	0.93	0.93	0.93	0.92 ↓	0.89	0.90 ↑	0.89	0.92 ↑
		Intersectional metric	0.63	0.65 ↑	0.63	0.18 ↓	0.74	0.70 ↓	0.74	0.26 ↓	0.63	0.65 ↑	0.63	0.17
	RF	Sex	0.56	0.50 ↓	0.56	0.92 ↑	0.71	0.68 ↓	0.71	0.98 ↑	0.56	0.50 ↓	0.56	0.92 ↑
		Race	0.87	0.82 ↓	0.87	0.80 ↓	0.85	0.84 ↓	0.85	0.78 ↓	0.87	0.81 ↓	0.87	0.80 ↓
		Ethnicity	0.88	0.85 ↓	0.88	0.85 ↓	0.91	0.89 ↓	0.91	0.89 ↓	0.88	0.85 ↓	0.88	0.85 ↓
		Intersectional metric	0.35	0.55 ↑	0.35	0.18 ↓	0.40	0.54 ↑	0.40	0.27 ↓	0.34	0.55 ↑	0.34	0.17
Respiratory	LR	Sex	0.68	0.70 ↑	0.68	0.96 ↑	0.82	0.83 ↑	0.82	0.95 ↑	0.68	0.69 ↑	0.68	0.96 ↑
		Race	0.64	0.62 ↓	0.64	0.84 ↑	0.76	0.75 ↓	0.76	0.90 ↑	0.64	0.62 ↓	0.64	0.84 ↑
		Ethnicity	0.66	0.67 ↑	0.66	0.75 ↑	0.72	0.73 ↑	0.72	0.78 ↑	0.66	0.67 ↑	0.66	0.75 ↑
		Intersectional metric	0.37	0.38 ↑	0.37	0.12 ↓	0.66	0.66	0.67 ↑	0.49 ↓	0.36	0.38 ↑	0.37 ↑	0.12
	RF	Sex	0.96	0.97 ↑	0.96	0.98 ↑	0.95	0.96 ↑	0.95	0.97 ↑	0.96	0.97 ↑	0.96	0.98 ↑
		Race	0.76	0.76	0.76	0.80 ↑	0.85	0.85	0.85	0.87 ↑	0.76	0.76	0.76	0.80 ↑
		Ethnicity	0.80	0.81 ↑	0.80	0.85 ↑	0.79	0.80 ↑	0.79	0.83 ↑	0.81	0.81	0.81	0.85 ↑
		Intersectional metric	0.18	0.18	0.18	0.14 ↓	0.58	0.58	0.58	0.57 ↓	0.17	0.17	0.17	0.14

Source Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Notes: Results are shown as the average across 100 bootstrap samples for each model specifications after optimizing the combination of all protected attributes. We specified the protected attribute as the second largest non-missing category for the demographic categories sex (male), race (Black or African American), and ethnicity (Hispanic or Latino). The Baseline columns represent model fairness before any bias mitigation techniques are applied.

Table A.4. (*continued*)

↑ increase from baseline

↓ decrease from baseline

AUROC = area under the receiver operating characteristic curve; LR = Logistic Regression; MAAT = mitigating algorithmic bias with adversarial training; NHPI = Native Hawaiian or Pacific Islander; PRAUC = area under the precision-recall curve; RF = Random Forest

Table A.5. Subgroup performance for test data before and after bias mitigation when optimizing multiple attributes

Symptom Cluster	Model Type	Demographic Subgroup	AUROC				PRAUC			
			Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
Cognitive	LR	All	0.73	0.72 ↓	0.72 ↓	0.73	0.36	0.36	0.36	0.36
		Sex: Female	0.71	0.71	0.71	0.72 ↑	0.36	0.36	0.36	0.35 ↓
		Sex: Male	0.74	0.74	0.74	0.74	0.36	0.37 ↑	0.36	0.38 ↑
		Sex: Other	--	--	--	--	--	--	--	--
		Race: Asian	0.72	0.72	0.72	0.72	0.36	0.36	0.35 ↓	0.36
		Race: Black or African American	0.70	0.70	0.70	0.69 ↓	0.33	0.34 ↑	0.33	0.33
		Race: Missing or unknown	0.71	0.71	0.71	0.71	0.29	0.29	0.29	0.29
		Race: NHPI	--	--	--	--	--	--	--	--
		Race: Other	0.83	0.83	0.83	0.81 ↓	0.42	0.42	0.42	0.40 ↓
		Race: White	0.73	0.73	0.73	0.73	0.37	0.38 ↑	0.38 ↑	0.37
		Ethnicity: Hispanic or Latino	0.73	0.74 ↑	0.74 ↑	0.74 ↑	0.32	0.33 ↑	0.32	0.33 ↑
		Ethnicity: Missing or unknown	0.73	0.73	0.73	0.73	0.32	0.33 ↑	0.32	0.33 ↑
		Ethnicity: Not Hispanic or Latino	0.72	0.72	0.72	0.72	0.37	0.37	0.37	0.36 ↓
	RF	All	0.72	0.72	0.72	0.72	0.37	0.37	0.37	0.37
		Sex: Female	0.72	0.72	0.72	0.72	0.37	0.37	0.37	0.37
		Sex: Male	0.73	0.73	0.73	0.73	0.38	0.38	0.38	0.38
		Sex: Other	--	--	--	--	--	--	--	--
		Race: Asian	0.74	0.74	0.74	0.74	0.38	0.38	0.38	0.38
		Race: Black or African American	0.68	0.68	0.68	0.68	0.34	0.35 ↑	0.34	0.34
		Race: Missing or unknown	0.74	0.74	0.74	0.73 ↓	0.34	0.35 ↑	0.34	0.33 ↓
		Race: NHPI	--	--	--	--	--	--	--	--
		Race: Other	0.83	0.83	0.83	0.83	0.42	0.42	0.42	0.42
		Race: White	0.72	0.72	0.72	0.72	0.38	0.38	0.38	0.38
		Ethnicity: Hispanic or Latino	0.73	0.74 ↑	0.73	0.73	0.34	0.35 ↑	0.34	0.34
		Ethnicity: Missing or unknown	0.76	0.76	0.76	0.76	0.36	0.36	0.36	0.36
		Ethnicity: Not Hispanic or Latino	0.72	0.72	0.72	0.72	0.38	0.38	0.38	0.38

Table A.5. (*continued*)

Symptom Cluster	Model Type	Demographic Subgroup	AUROC				PRAUC			
			Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
Fatigue	LR	All	0.60	0.60	0.60	0.59 ↓	0.35	0.38 ↑	0.35	0.29 ↓
		Sex: Female	0.57	0.56 ↓	0.57	0.58 ↑	0.43	0.45 ↑	0.43	0.30 ↓
		Sex: Male	0.57	0.58 ↑	0.57	0.60 ↑	0.17	0.21 ↑	0.17	0.27 ↑
		Sex: Other	0.56	0.55 ↓	0.56	0.61 ↑	0.26	0.27 ↑	0.27 ↑	0.25 ↓
		Race: Asian	0.57	0.58 ↑	0.58 ↑	0.55 ↓	0.21	0.22 ↑	0.21	0.16 ↓
		Race: Black or African American	0.62	0.62	0.62	0.61 ↓	0.35	0.42 ↑	0.35	0.35
		Race: Missing or unknown	0.60	0.61 ↑	0.60	0.59 ↓	0.23	0.29 ↑	0.23	0.22 ↓
		Race: NHPI	0.56	0.56	0.57 ↑	0.55 ↓	0.19	0.20 ↑	0.20 ↑	0.17 ↓
		Race: Other	0.62	0.63 ↑	0.63 ↑	0.58 ↓	0.34	0.36 ↑	0.34	0.30 ↓
		Race: White	0.59	0.59	0.59	0.58 ↓	0.37	0.39 ↑	0.37	0.29 ↓
		Ethnicity: Hispanic or Latino	0.60	0.60	0.60	0.58 ↓	0.33	0.40 ↑	0.33	0.27 ↓
		Ethnicity: Missing or unknown	0.59	0.60 ↑	0.59	0.58 ↓	0.20	0.21 ↑	0.20	0.19 ↓
		Ethnicity: Not Hispanic or Latino	0.60	0.60	0.60	0.59 ↓	0.37	0.39 ↑	0.37	0.30 ↓
	RF	All	0.60	0.60	0.60	0.60	0.33	0.37 ↑	0.33	0.30 ↓
		Sex: Female	0.59	0.57 ↓	0.59	0.59	0.36	0.41 ↑	0.36	0.31 ↓
		Sex: Male	0.60	0.61 ↑	0.60	0.61 ↑	0.26	0.29 ↑	0.26	0.30 ↑
		Sex: Other	0.89	0.85 ↓	0.89	0.75 ↓	0.54	0.53 ↓	0.54	0.36 ↓
		Race: Asian	0.57	0.59 ↑	0.57	0.57	0.25	0.29 ↑	0.25	0.24 ↓
		Race: Black or African American	0.62	0.61 ↓	0.62	0.62	0.37	0.42 ↑	0.37	0.36 ↓
		Race: Missing or unknown	0.62	0.63 ↑	0.62	0.61 ↓	0.27	0.33 ↑	0.27	0.26 ↓
		Race: NHPI	0.55	0.56 ↑	0.55	0.54 ↓	0.22	0.27 ↑	0.22	0.21 ↓
		Race: Other	0.60	0.64 ↑	0.60	0.59 ↓	0.29	0.36 ↑	0.29	0.25 ↓
		Race: White	0.59	0.59	0.59	0.59	0.33	0.37 ↑	0.33	0.30 ↓
		Ethnicity: Hispanic or Latino	0.60	0.60	0.60	0.59 ↓	0.31	0.40 ↑	0.31	0.28 ↓
		Ethnicity: Missing or unknown	0.62	0.63 ↑	0.62	0.62	0.25	0.28 ↑	0.25	0.26 ↑
		Ethnicity: Not Hispanic or Latino	0.60	0.60	0.60	0.59 ↓	0.34	0.37 ↑	0.34	0.31 ↓

Table A.5. *(continued)*

Symptom Cluster	Model Type	Demographic Subgroup	AUROC				PRAUC			
			Baseline	Reweighting	MAAT	FairMask	Baseline	Reweighting	MAAT	FairMask
Respiratory	LR	All	0.61	0.61	0.61	0.61	0.29	0.29	0.29	0.28 ↓
		Sex: Female	0.61	0.60 ↓	0.60 ↓	0.61	0.31	0.31	0.31	0.28 ↓
		Sex: Male	0.61	0.61	0.61	0.61	0.26	0.26	0.26	0.29 ↑
		Sex: Other	--	--	--	--	--	--	--	--
		Race: Asian	0.60	0.60	0.60	0.60	0.26	0.26	0.26	0.26
		Race: Black or African American	0.60	0.60	0.60	0.61 ↑	0.35	0.36 ↑	0.36 ↑	0.31 ↓
		Race: Missing or Unknown	0.60	0.60	0.60	0.60	0.20	0.21 ↑	0.20	0.20
		Race: NHPI	0.62	0.62	0.63 ↑	0.62	0.23	0.24 ↑	0.25 ↑	0.24 ↑
		Race: Other	0.58	0.58	0.58	0.58	0.31	0.31	0.31	0.30 ↓
		Race: White	0.60	0.61 ↑	0.60	0.61 ↑	0.29	0.29	0.29	0.29
		Ethnicity: Hispanic or Latino	0.59	0.59	0.59	0.59	0.22	0.23 ↑	0.22	0.23 ↑
		Ethnicity: Missing or unknown	0.60	0.60	0.60	0.59 ↓	0.18	0.18	0.18	0.18
		Ethnicity: Not Hispanic or Latino	0.61	0.61	0.61	0.61	0.31	0.31	0.31	0.30 ↓
	RF	All	0.62	0.62	0.62	0.62	0.30	0.30	0.30	0.31 ↑
		Sex: Female	0.62	0.62	0.62	0.62	0.31	0.31	0.31	0.31
		Sex: Male	0.61	0.61	0.61	0.62 ↑	0.29	0.29	0.29	0.30 ↑
		Sex: Other	--	--	--	--	--	--	--	--
		Race: Asian	0.61	0.61	0.61	0.61	0.28	0.28	0.28	0.28
		Race: Black or African American	0.62	0.62	0.62	0.62	0.34	0.35 ↑	0.34	0.34
		Race: Missing or unknown	0.62	0.62	0.62	0.62	0.26	0.26	0.26	0.26
		Race: NHPI	0.61	0.61	0.61	0.62 ↑	0.25	0.25	0.25	0.27 ↑
		Race: Other	0.57	0.57	0.57	0.58 ↑	0.26	0.26	0.26	0.28 ↑
		Race: White	0.62	0.62	0.62	0.62	0.30	0.30	0.30	0.31 ↑
		Ethnicity: Hispanic or Latino	0.59	0.59	0.59	0.60 ↑	0.25	0.25	0.25	0.26 ↑
		Ethnicity: Missing or unknown	0.62	0.62	0.62	0.62	0.25	0.25	0.25	0.25
		Ethnicity: Not Hispanic or Latino	0.62	0.62	0.62	0.62	0.31	0.32 ↑	0.31	0.32 ↑

Source Analysis completed in the National COVID Cohort Collaborative Data Enclave.

Table A.5. (*continued*)

Note: Results are shown as the average across 100 bootstrap samples for each model specifications after optimizing the combination of all protected attributes. Bias mitigation was performed via reweighting, MAAT, and FairMask based on the protected attribute. We specified the protected attribute as the second largest non-missing category for the demographic categories sex (male), race (Black or African American), and ethnicity (Hispanic or Latino). The demographic subgroups are shown in accordance with the demographic characteristics on which bias mitigation was performed. "All" is the AUROC and PRAUC for the model across characteristic subgroups. The Baseline columns represent model performance before any bias mitigation techniques are applied.

↑ increase from baseline

↓ decrease from baseline

-- We were unable to calculate the AUROC or PRAUC for some demographic characteristic subgroups due to small sample size.

PRAUC: area under the precision-recall curve; AUROC: Area under the receiver operating characteristic curve; NHPI: Native Hawaiian or Pacific Islander