Disclaimer: all content falls under 'views expressed' and does not reflect my employer

Harnessing AI for Health Equity and Inclusivity in Risk-Based Monitoring

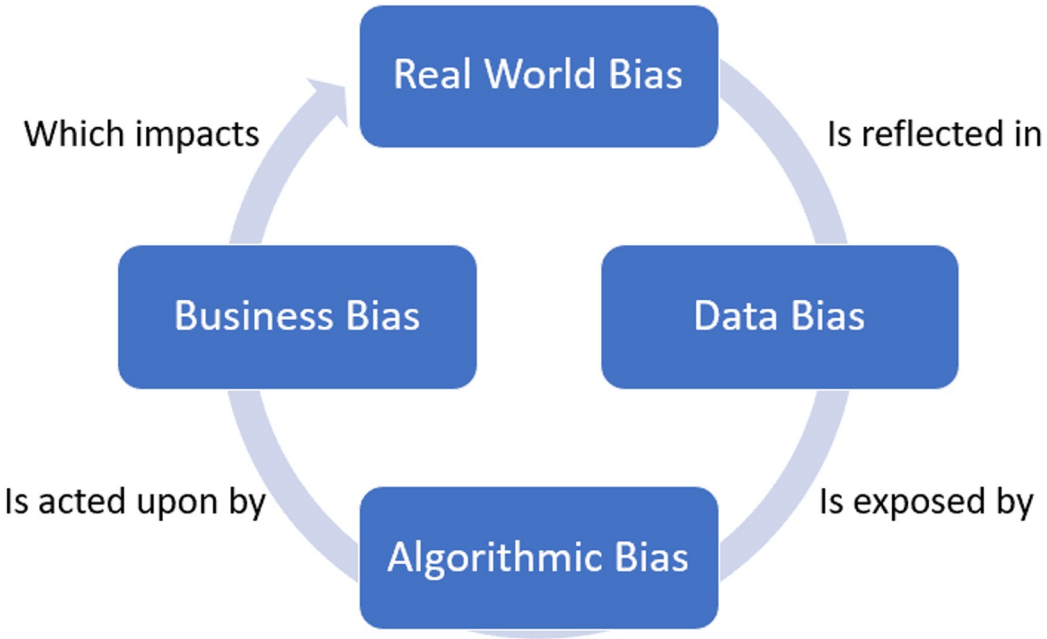
PHUSE Connect Innovation Challenge

Finn Janson, Data Scientist

25th February 2024

Table of contents

Racial Disparities in Trials


Real World Bias
Is reflected in
Data Bias
Is exposed by
Algorithmic Bias
Is acted upon by
Business Bias
Which impacts

Compared to White Patients, Black Patients have:

40% Higher Breast Cancer Mortality (women)

With 5% representation in clinical trials (2020)

Worse MS Prognosis

With 1% representation in clinical trials (2020)

22% Higher Mortality from Melanoma

With massive underrepresentation in public skin lesion images

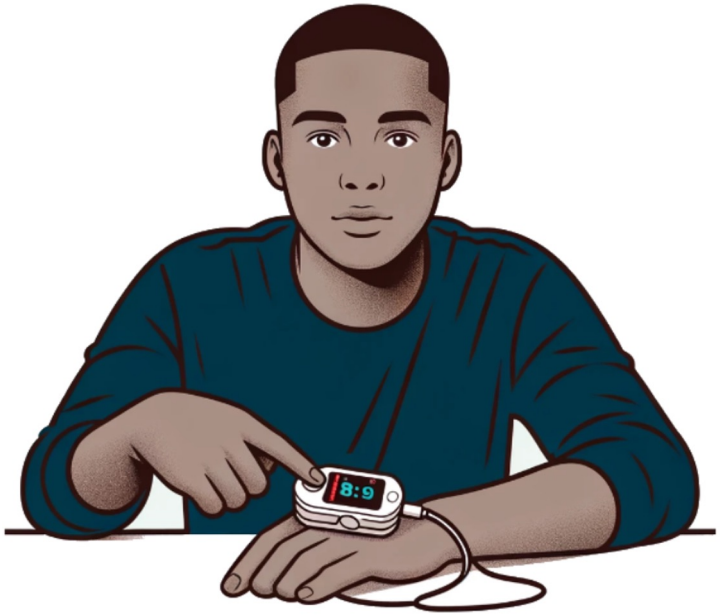
Bias in Pulse Oximeters

Used in

Respiratory and Cardiovascular Drug Trials

Medical Device Trials

Experimental Treatments (e.g., ARDS, TBI)



Provides

Non-Invasive Measurement of Oxygen Saturation (SpO2)

Gauge of Drug/Treatment Safety and Efficacy

Monitoring for Patients with Critical Oxygen Levels

Variation in skin tones skews the results for SpO2, leading to **underestimations of disease severity, misdiagnosis or even adverse events**

Error Rates Across Skin Tones

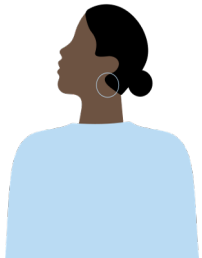
Represented by Size

Bias in Measurements

Darker-skinned patients often 3x as likely to have hypoxemia missed by pulse oximeters

Even a small bias of 1% overestimating SaO2 can spike hidden hypoxemia

Tendency to overestimate oxygen saturation by up to 4%, worsening as patient SaO2 levels drop



Algorithmic Fairness

Necessitates fairness checkpoints when considering risk-based strategies

Bias in Estimations

Skewed results for efficacy, estimand effect and endpoints

Unreliable Quality of Life assessment

Misinformed Oxygen treatment

Data Integrity

Necessitates recalculation based on more reliable data points or adjusted values

Data Integrity

Risk-based Monitoring Approach

Algorithmic Fairness

Implications for RBQM

Endpoints

SpO2 usage in endpoints must address data quality issue

Bias in Pulse Oximeter readings (SpO2) has major implications from a data quality and patient safety perspective

Quality Tolerance Limits

Set to limit the SpO2 readings informed by historical data and experts

Estimands

Altering oxygen levels based on SpO2 readings can lead to adverse outcomes

Intercurrent Events

Intercurrent event (ICEs) since it affects interpretation of results

Source Data Review

SDR of SpO2 data to ensure data quality of SpO2 readings

Endpoints

SpO2 usage in endpoints must address data quality issue

Bias in Pulse Oximeter readings (SpO2) has major implications from a data quality and patient safety perspective

Quality Tolerance Limits

Set to limit the SpO2 readings informed by historical data and experts

Estimands

Altering oxygen levels based on SpO2 readings can lead to adverse outcomes

Intercurrent Events

Intercurrent event (ICEs) since it affects interpretation of results

Source Data Review

SDR of SpO2 data to ensure data quality of SpO2 readings

Endpoints

SpO2 usage in endpoints must address data quality issue

Bias in Pulse Oximeter readings (SpO2) has major implications from a data quality and patient safety perspective

Quality Tolerance Limits

Set to limit the SpO2 readings informed by historical data and experts

Estimands

Altering oxygen levels based on SpO2 readings can lead to adverse outcomes

Intercurrent Events

Intercurrent event (ICEs) since it affects interpretation of results

Source Data Review

SDR of SpO2 data to ensure data quality of SpO2 readings

Endpoints

SpO2 usage in endpoints must address data quality issue

Bias in Pulse Oximeter readings (SpO2) has major implications from a data quality and patient safety perspective

Quality Tolerance Limits

Set to limit the SpO2 readings informed by historical data and experts

Estimands

Altering oxygen levels based on SpO2 readings can lead to adverse outcomes

Intercurrent Events

Intercurrent event (ICEs) since it affects interpretation of results

Source Data Review

SDR of SpO2 data to ensure data quality of SpO2 readings

Endpoints

SpO2 usage in endpoints must address data quality issue

Bias in Pulse Oximeter readings (SpO2) has major implications from a data quality and patient safety perspective

Quality Tolerance Limits

Set to limit the SpO2 readings informed by historical data and experts

Estimands

Altering oxygen levels based on SpO2 readings can lead to adverse outcomes

Intercurrent Events

Intercurrent event (ICEs) since it affects interpretation of results

Source Data Review

SDR of SpO2 data to ensure data quality of SpO2 readings

Using Machine Learning for QTLs

Estimation of QTLs

Simple statistical models or Bayesian approaches

Deciding QTLs for SpO2 measurements

Alternative Data Sources

QTLs could be estimated from real world data in lieu of sufficient trial data

Limitations of QTL setting

However, setting QTL thresholds based only on past evidence, we reinforce exclusion.

Estimation of QTLs

Simple statistical models or Bayesian approaches

Deciding QTLs for SpO2 measurements

Alternative Data Sources

QTLs could be estimated from real world data in lieu of sufficient trial data

Limitations of QTL setting

However, setting QTL thresholds based only on past evidence, we reinforce exclusion.

Estimation of QTLs

Simple statistical models or Bayesian approaches

Deciding QTLs for SpO2 measurements

Alternative Data Sources

QTLs could be estimated from real world data in lieu of sufficient trial data

Limitations of QTL setting

However, setting QTL thresholds based only on past evidence, we reinforce exclusion.

Estimation of QTLs

Simple statistical models or Bayesian approaches

Deciding QTLs for SpO2 measurements

Alternative Data Sources

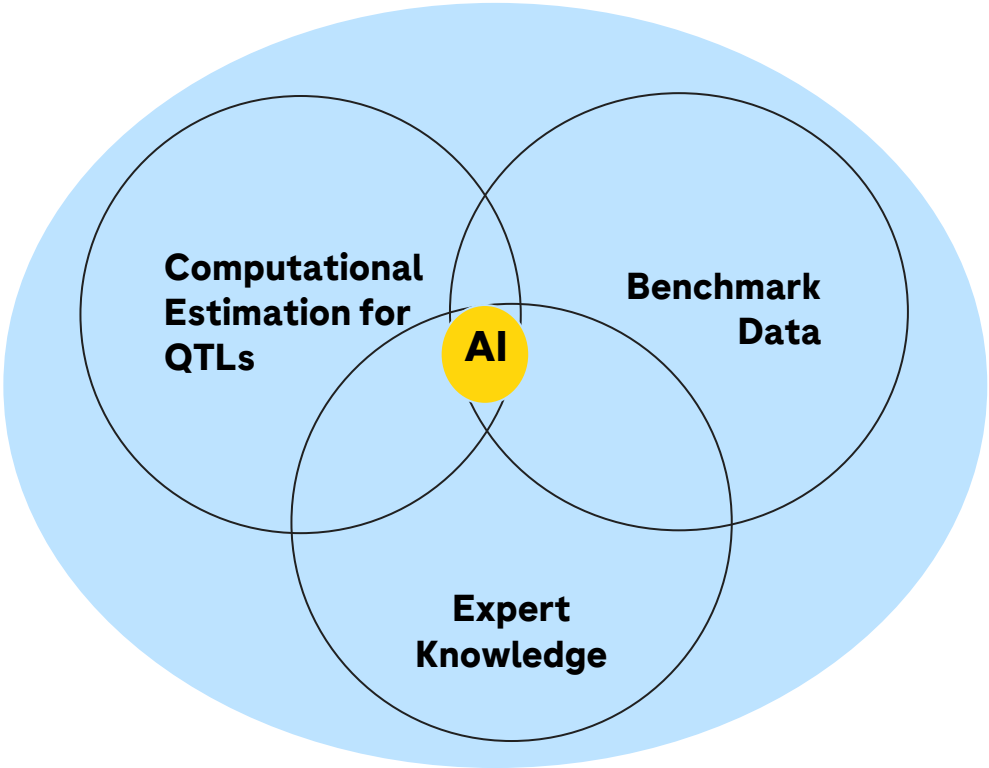
QTLs could be estimated from real world data in lieu of sufficient trial data

Limitations of QTL setting

However, setting QTL thresholds based only on past evidence, we reinforce exclusion.

Impact of biased readings on Trials

Subtitle goes here but is not mandatory



Historical clinical data suffers from systematic disparities leading to underrepresentation of marginalized groups

However, a ML model trained on SaO2 levels with respect to fairness can mitigate this

In this case, **the ML model represents less bias** (over or underprediction) of SaO2 levels than pulse oximeter readings, as well as **signalling patient groups with high measurement error** from device

Proposal of Model

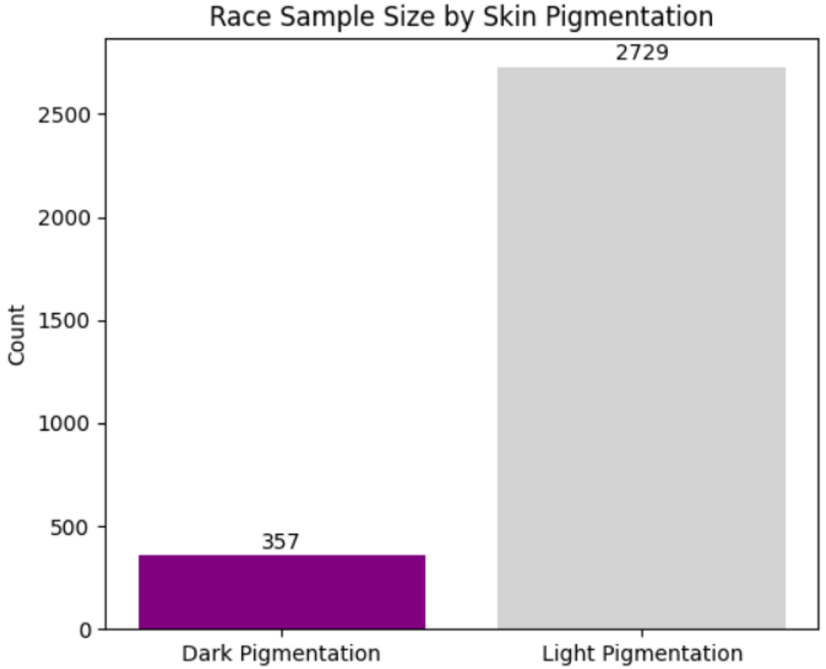
Aim

Addressing biased pulse oximeter readings across demographics by applying machine learning to improve oxygen saturation (SaO2) prediction

Cohort

The model utilizes the large-scale MIMIC critical care database containing ***over 50,000 patients*** *admitted from 2008-2019.*

After train-test split, the training set comprised ***12,202 samples****, while the test set contained 3,086 samples, divided into 5 k-folds for hyperparameter optimization*



Dealing with implausible data

SaO2 (true oxygen saturation) and SpO2 (pulse oximeter value) outside 70-100% were excluded to remove implausible readings

Removed subjects with

- Over 5-minute lag between readings
- Over 50% missing data
- Unclear racial classification or extremely rare race

Preparing features

After categorizing patients by inferred skin pigmentation, imputing missing values with mean or mode, features were narrowed down to 16 based on performance, clinical relationship with SaO2 and and collinearity

Selected features were related to patient status across Respiratory, Cardiovascular, Laboratory, Clinical Status, and Demographics and Treatment information

Model Selection

CatBoostRegressor, ideal for **handling high-cardinality categories such as race, uses efficient encoding to boost performance**

Key Components

- Employs ordered boosting for regularization
- Growing trees level-wise to avoid overfitting
- Fast training and inference, even with large datasets, making it suitable for medical device integration.

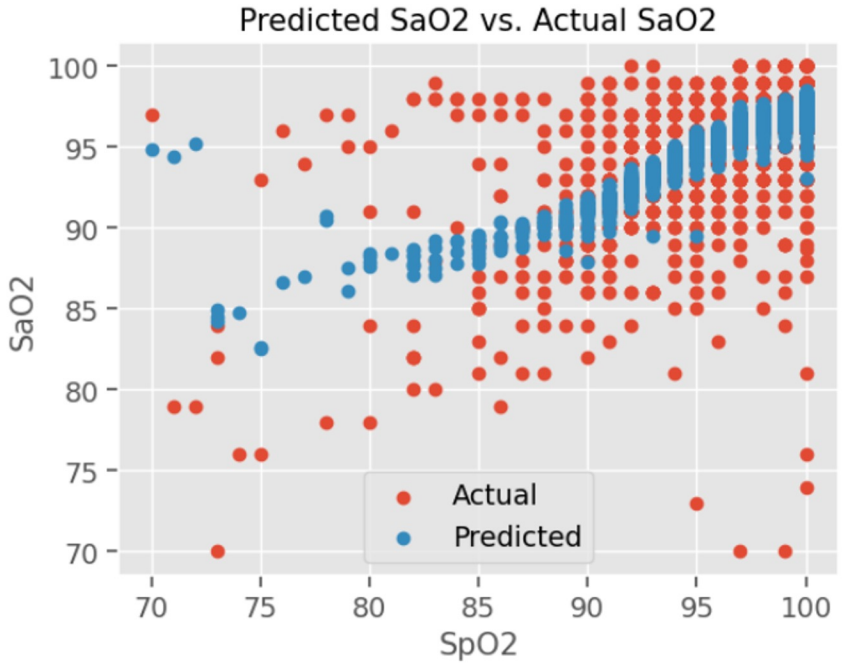
Fairness as Objective Function

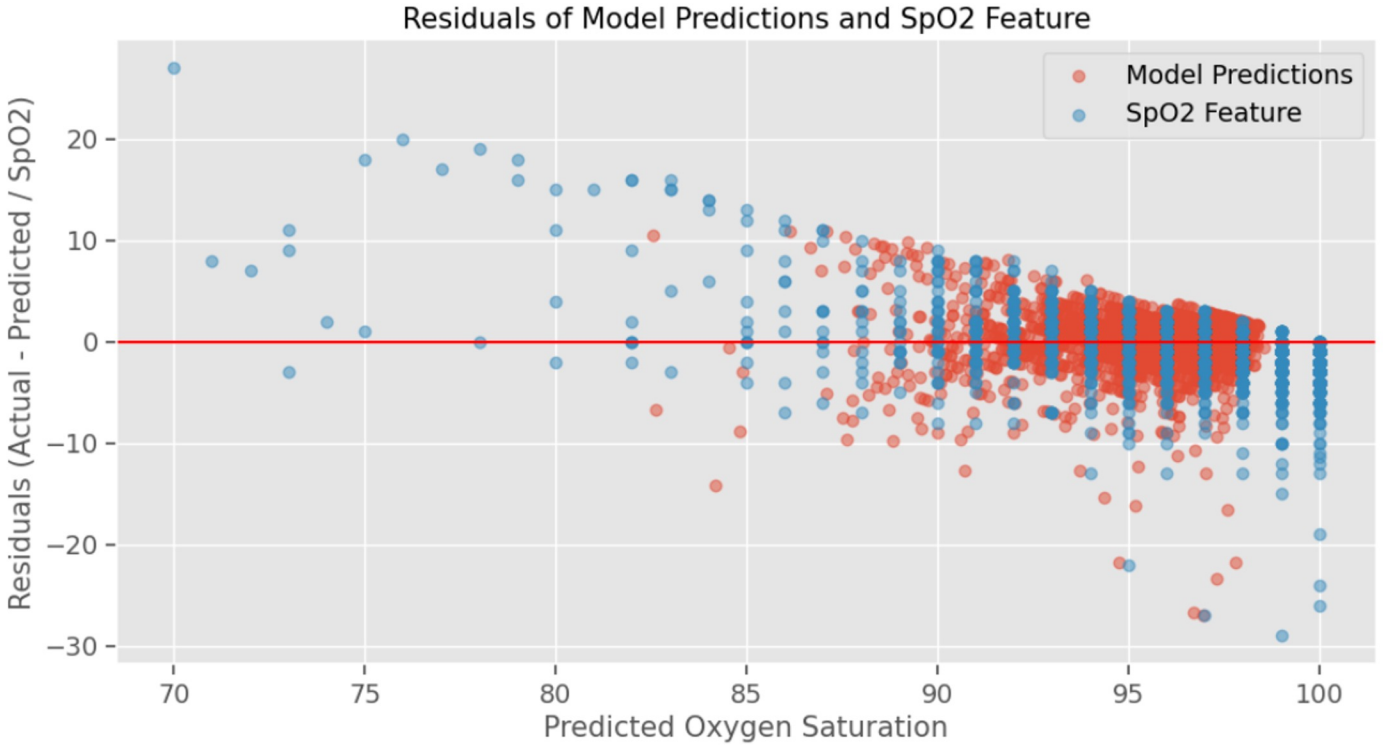
To ensure equitable model performance across races, model is optimized for *weighted MSE* (mean squared error) for patients of dark skin pigmentation

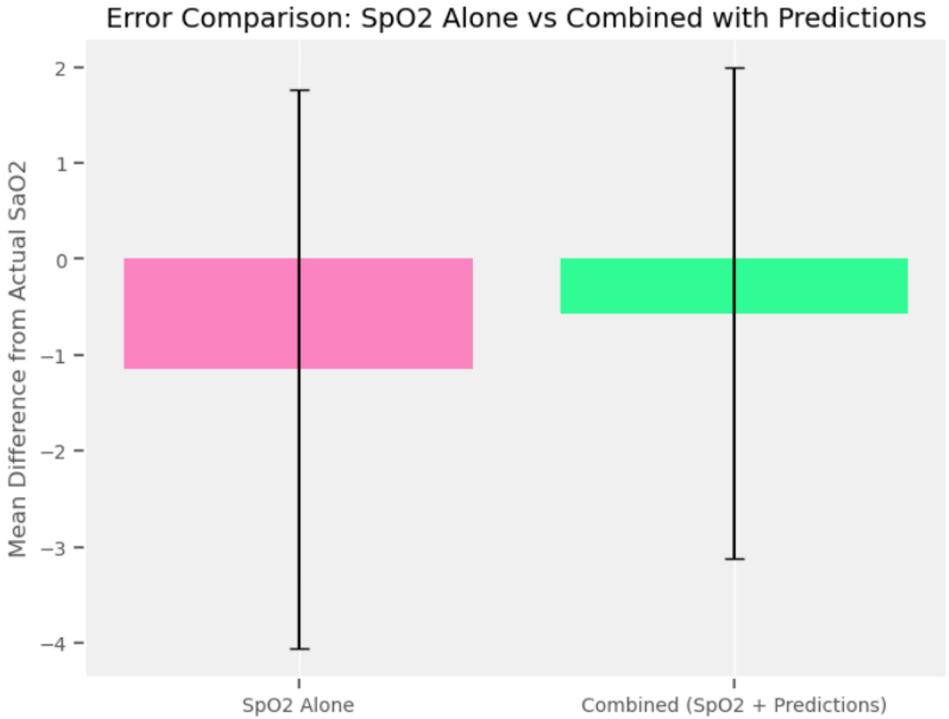
Hyperparameter Selection

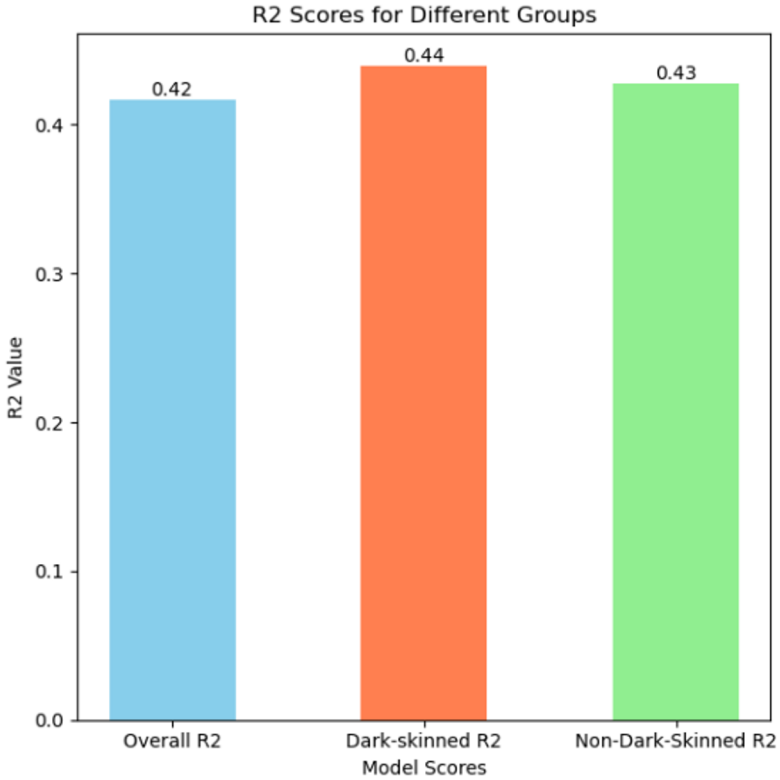
Bayesian optimization was chosen for its time-cost efficiency. This was used to find best values for learning rate, max depth, subsampling and number of estimator

Results of Model


Predicted SaO2 vs. Actual SaO2
SaO2
SpO2
100
95
90
85
80
75
70
70
75
80
85
90
95
100
Actual
Predicted

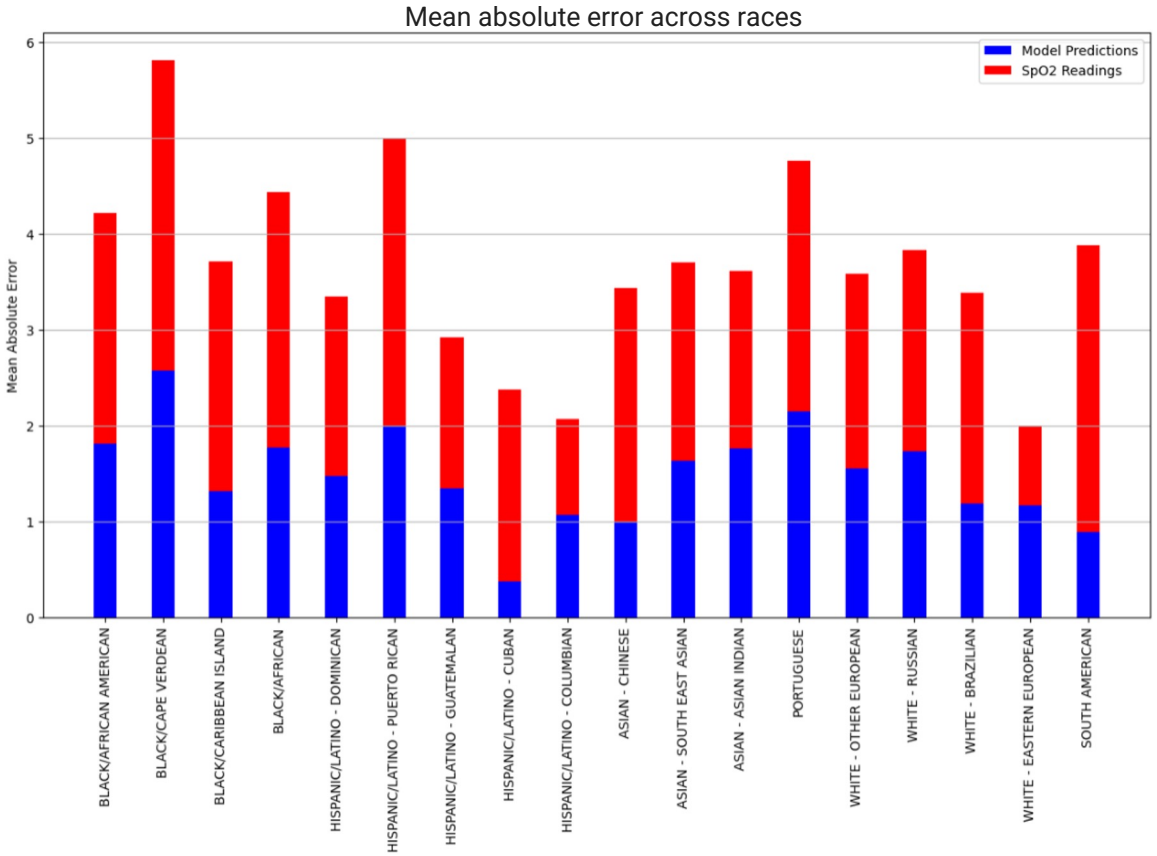

Residuals of Model Predictions and SpO2 Feature
Model Predictions
SpO2 Feature
Residuals (Actual - Predicted / SpO2)
20
10
0
−10
−20
−30
70
75
80
85
90
95
100
Predicted Oxygen Saturation

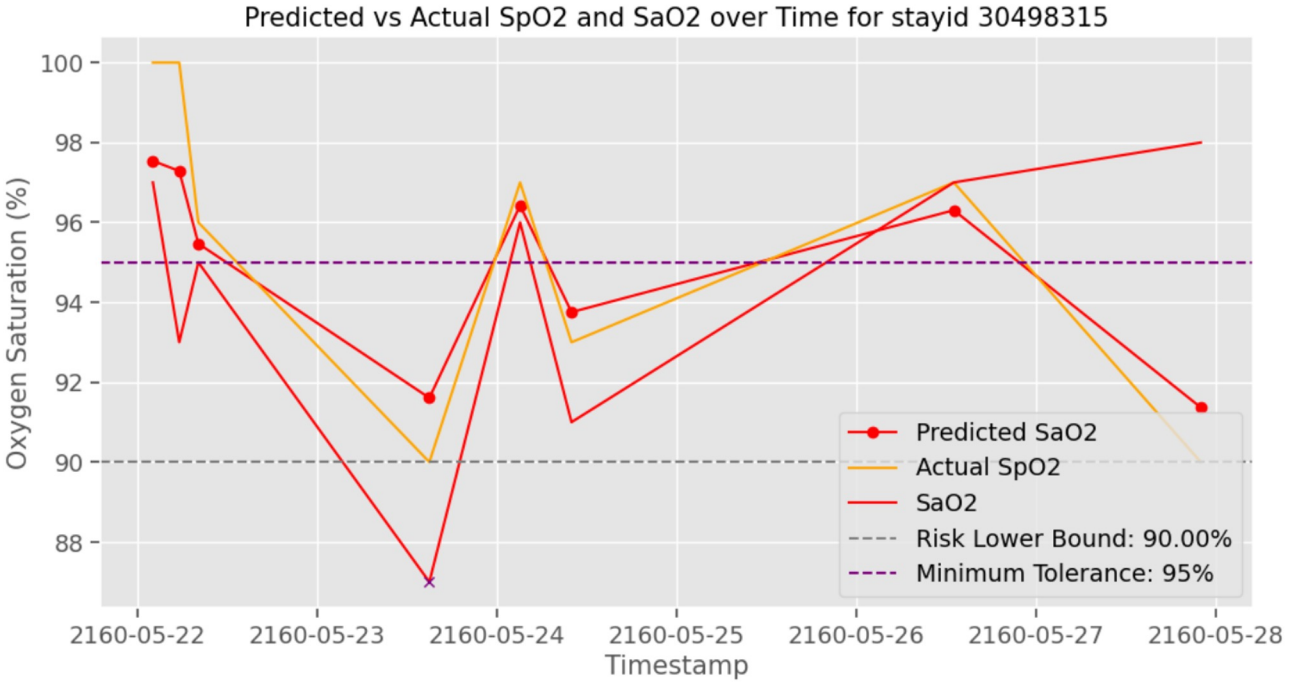

Error Comparison: SpO2 Alone vs Combined with Predictions
Mean Difference from Actual SaO2
2
1
0
−1
−2
−3
−4
SpO2 Alone
Combined (SpO2 + Predictions)



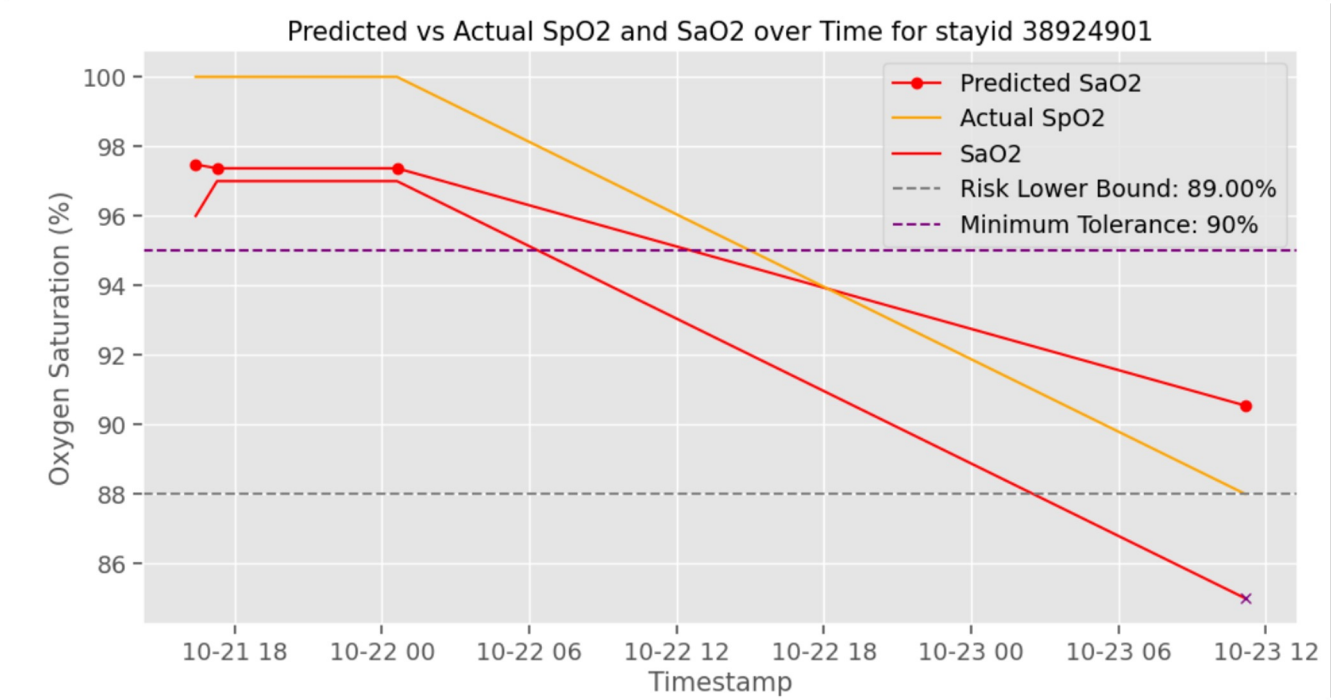
The R2 values indicates that the model explains 44% of the variability for the dark-skinned subgroup

Crucially, the R2 values are not lower for dark-skin patients than they are for non-dark skinned despite class imbalance (2729 of patients in the training data set with no skin pigmentation and 357 with skin pigmentation)

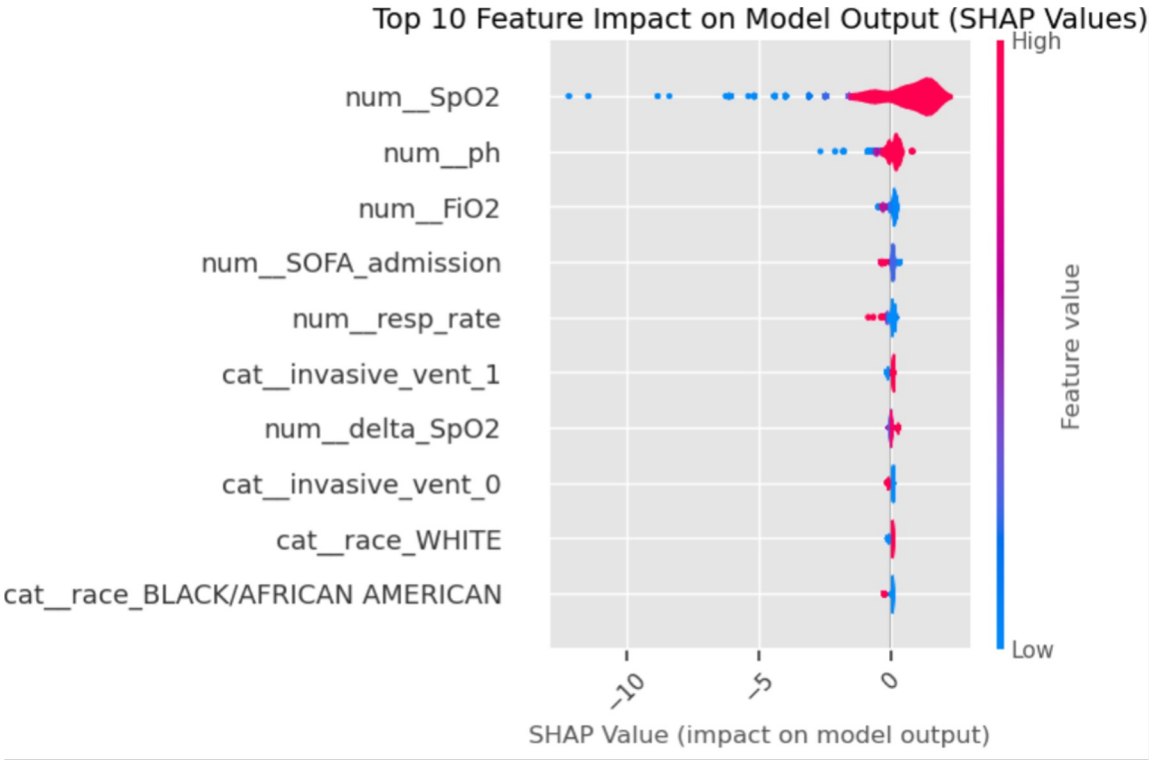

Mean absolute error across races
Model Predictions
SpO2 Readings
Mean Absolute Error
0
1
2
3
4
5
6
BLACK/AFRICAN AMERICAN
BLACK/CAPE VERDEAN
BLACK/CARIBBEAN ISLAND
BLACK/AFRICAN
HISPANIC/LATINO - DOMINICAN
HISPANIC/LATINO - PUERTO RICAN
HISPANIC/LATINO - GUATEMALAN
HISPANIC/LATINO - CUBAN
HISPANIC/LATINO - COLUMBIAN
ASIAN - CHINESE
ASIAN - SOUTH EAST ASIAN
ASIAN - ASIAN INDIAN
PORTUGUESE
WHITE - OTHER EUROPEAN
WHITE - RUSSIAN
WHITE - BRAZILIAN
WHITE - EASTERN EUROPEAN
SOUTH AMERICAN



x = hidden hypoxemia event



x = hidden hypoxemia event


Top 10 Feature Impact on Model Output (SHAP Values)
num__SpO2
num__ph
num__FiO2
num__SOFA_admission
num__resp_rate
cat__invasive_vent_1
num__delta_SpO2
cat__invasive_vent_0
cat__race_WHITE
cat__race_BLACK/AFRICAN AMERICAN
High
Feature value
Low
−10
−5
0
SHAP Value (impact on model output)

Future

Setting or informing **QTLs by values provided by ML models** trained on historical or benchmark datasets

In order to validate the readings, **true skin pigmentation must be captured** in a trial-specific dataset

The time-series element was not fully addressed: it will be necessary to evaluate model across **measurements from varied time points and frequencies**