Reconsidering the Algorithmic Fairness of Race Adjustment in Pulmonary Function Equations

Isha Puri [*], **Neil Sehgal** [*], **Usha Bhalla** [*]

[1]Harvard University

Abstract

Race correction in clinical medical algorithms is widely used across many medical sub-fields, including in cardiology, nephrology, pulmonology, and obstetric medicine. Recently, however, the value of race correction including its potential bias in spirometry tests has been called into question by medical experts [1, 2]. In this paper, we present the results of a study auditing the fairness of race in pulmonary function tests. We extract Forced expiratory volume in the first second (FEV1) spirometry values (which is useful to categorize the severity of obstructive lung diseases), demographic variables, and self-reported pulmonary and functional impairments for Black and White individuals from a national survey (n=5571). FEV1 reference values were calculated for each survey respondent using the Global Lung Initiative (GLI)-2012 White and Black equations. We assess the accuracy and fairness of the equation for each individual (1) using their corresponding race GLI equation, (2) using the White GLI equation for everyone, and (3) using an equation fit without race for everyone. To assess the accuracy of (3), we refit a race-neutral equation to a subsample of the respondents of the survey, and assess the accuracy of a race-conscious and race-neutral equation on a held out test set. Our results show that across symptoms, specificity is lowest for Blacks when using race-conscious equations; however, mathematical fairness is best for the current GLI-2012 equation with race taken into consideration. These results suggest that in contrast to some recent well publicized studies - both in medical literature [1] and in general media [2], further work is required for improving the fairness of spirometry test equations before the race adjustment factors can removed from widely practiced spirometry tests.

Introduction

The history of integrating race into medical treatment is a complicated one. Historically, the centuries-old idea of "racial essentialism" theorized that races were biologically distinct groups with distinct genetic makeups. Despite growing evidence against the belief that racial groups possess inherent biological differences, the idea's legacies remain and have been intertwined into countless clinical practices, diagnostic metrics, and health policies.

One such integration of race into clinical medical care includes the practice of race norming, in which the outputs of diagnostic algorithms are "adjusted" or "corrected" based on a patient's race. Race-normed equations are used in countless areas of medicine (cardiology, nephrology, obstetrics, urology, oncology, endocrinology, and pulmonology) and originated with 19th century studies concluding that Black people had lower lung capacities than White people [3].

Critics of the practice of race norming, however, suggest that propagating race-based medicine by using race-adjusted algorithms directs a disproportional number of resources towards White patients and forces Black people to show greater levels of suffering before receiving the same treatment. The practice gained widespread news coverage, for example, after the NFL stopped its use when determining eligibility for compensation after concussions. A widely discussed controversy that found that retired Black players were assumed to have lower baseline cognitive function levels than retired white players, and thus had to displayer higher levels of cognitive declines to receive the same financial awards [4].

The field's widespread disagreement begs the question - should we factor race into our clinical decisions?

In this paper, we investigate the use of race-norming in the field of pulmanory medicine, which is the branch of medicine that deals with the causes, diagnosis, prevention and treatment of diseases affecting the lungs. Pulmanologists use spirometry tests to assess how well a patient's lungs work by measuring how much air is inhaled, how much air is exhaled and how quickly a patient exhales. One of the results of spirometry tests is the FEV1 value, which represents the volume of breath exhaled with effort in one second [5]. These results from spirometry tests are used to diagnose asthma, chronic obstructive pulmonary disease (COPD) and other conditions that affect breathing.

In the US, spirometry tests use race-specific reference values or "correction factors" based on results derived from data collected by the Global Lung Initiative in 2012. This data was collected from 72,031 non-smoking healthy individuals between the ages of 3 and 95 (Caucasians (n=57,395), African–Americans (n=3,545), and North (n=4,992) and South East Asians (n=8,255))[6], and the equations provide "Lower Limit of Normal" values for spirometry based on one's sex, age, height, and race. Studies

[*]These authors contributed equally.

have shown that these race based correction factors require Black people to show higher levels of lung function decline to be considered for the same disability resources or medical treatment as their white counterparts [3]. There is also evidence that using race-based correction factors does *not* help predict chronic lower respiratory disease events [7]. Such results all suggest that differences in race don't reflect biology as much as they do the effects that systematic racism has had on health socially, including differences in air quality, nutrition, etc. It follows that using race-normed medicine normalizes poor lung health and propagates racial health inequities - instead of normalizing inherent differences in lung capaciity, these eqautions might just be *normalizing structural inequities in healthcare and society*.

In this paper, we aim to analyze the fairness of race-normed spirometry equations. We also aim to discuss the results of removing race from such equations entirely.

Related Works

Previous work provides some preliminary results regarding the use of race adjustment in lung function spirometry tests. For instance, Ekstrom and Mannino assessed the effect of the GLI-2012's race-specific reference equations in relation to prediction of breathlessness and mortality. To understand the effect of a race-neutral equation, they investigate the use of the GLI-2012 White-reference equations on Blacks compared to Black-reference equations, finding that race-specific reference values does not improve prediction [8]. In this work, we take an algorithmic fairness approach to understanding race-adjustment in lung function tests. In addition, we not only investigate the use of White-reference values for Blacks, but also the use of a truly race-neutral equation.

Methods

To understand the impact of race correction in pulmonary function tests, we extracted data from a national survey, and analyzed the sensitivity, specificity, and positive and negative predictive values using different race correction specifications. Our three different specifications are as follows:

1. As a baseline, we consider the fairness of the existing GLI equation fit to different racial groups and tested on those corresponding groups.
2. We then analyze the efficacy of using the equation trained on Whites and tested on all populations during inference, as was suggested by [8].
3. Finally, we consider the equation fit to both Whites and Blacks during training and tested on both populations as well.

Data

The present study is a cross-sectional analysis of data from the 2007-2012 National Health and Nutrition Examination Survey (NHANES). The data in this study include demographic data on height, age, gender, and race; FEV1 spirometry values; self-reported smoking behavior (Smoked at least 100 cigarettes in life); and a range of self-reported pulmonary and physical functional impairments including breathlessness, wheezing, dry cough, physical work limitations, and general health.

Primary analysis was restricted to non-Hispanic Black and non-Hispanic White respondents aged 18 or older. Of the 18,359 White or Black survey respondents, 6,145 were excluded for missing data on race, gender, height, age, or FEV1. An additional 3,156 respondents were excluded for being under the age of 18, leaving a total sample of 9,058 individuals. Table 1 displays summary statistics.

Table 1: Summary Statistics. Values are mean with standard deviation shown in parentheses.

	Overall	**Black**	**White**
	9058	3130	5928
Age	46.4 (17.4)	45.3 (17.4)	47.0 (17.3)
Height	1.7 (0.1)	1.7 (0.1)	1.7 (0.1)
Gender			
Female	4484 (49.5)	1569 (50.1)	2915 (49.2)
Male	4574 (50.5)	1561 (49.9)	3013 (50.8)
FEV1	3078.4 (938.0)	2796.2 (838.5)	3227.4 (953.4)

Prevalence of pulmonary and physical functional impairments and rates of missing values are listed in table 2.

Table 2: Impairments. Percentages shown in parentheses.

		Overall	**Black**	**White**
Breathlessness	0	3714 (41.0)	1276 (40.8)	2438 (41.1)
	1	1853 (20.5)	619 (19.8)	1234 (20.8)
	-	3491 (38.5)	1235 (39.5)	2256 (38.1)
Wheezing	0	7673 (84.7)	2672 (85.4)	5001 (84.4)
	1	1377 (15.2)	457 (14.6)	920 (15.5)
	-	8 (0.1)	1 (0.0)	7 (0.1)
Dry Cough	0	8578 (94.7)	2976 (95.1)	5602 (94.5)
	1	475 (5.2)	154 (4.9)	321 (5.4)
	-	5 (0.1)		5 (0.1)
Limited in Work	0	6971 (77.0)	2363 (75.5)	4608 (77.7)
	1	1640 (18.1)	550 (17.6)	1090 (18.4)
	-	447 (4.9)	217 (6.9)	230 (3.9)
Bad Health	0	7021 (77.5)	2185 (69.8)	4836 (81.6)
	1	1462 (16.1)	689 (22.0)	773 (13.0)
	-	575 (6.3)	256 (8.2)	319 (5.4)

Fairness

To mathematically assess the fairness of race adjustment, we consider predictive rate parity, equalized opportunity, and specificity parity. These metrics consider the parity of the four most relevant accuracy metrics in clinical studies (sensitivity, specificity, PPV, and NPV). Predictive rate parity is met if both PPV and NPV are equal between Blacks and White, equalized opportunity is met if TPR is equal between groups, and specificity parity is met if TNR is equal between groups.

We determine if a specific parity is met by whether or not the p-value of a chi-squared test between the two racial groups is greater than 0.05.

Refitting GLI-2012

In order to study how fitting the GLI equation to each racial group affects accuracy and fairness for Blacks and Whites, we fit a quantile regression model with $q = 0.05$ on the same features as the GLI equation to predict the corresponding LLN of FEV_1 given someone's age, gender, height, and potentially race [9]. Models were fit on data from the NHANES 2007-2012 surveys, as the data from the GLI equation is not publicly available. The current GLI equations and quantile regressions trained on individual racial populations yield similar results for both Blacks and Whites, verifying the use of quantile regression as a proxy. We compare models trained on non-smoking White and Black respondents and tested on a held-out sample of White or Black people. To increase the statistical power of our tests, we perform k-fold cross validation with $k = 5$ and compound test results before statistical analysis.

Results

Results for the performance of the three different equation specifications, five different outcomes, and four metrics are shown below.

The results in Table 3 conclude that the original GLI-2012 equation, which is trained implicitly on race, is unfair for specificity and NPV across most outcomes, and has mixed fairness for sensitivity and PPV depending on the outcome. Rows with p-values less than 0.05 are considered unfair for that metric and outcome. Note that high sensitivity ensures that patients with poor lung function actually receive the care they deserve, and that sensitivity for Blacks is always lower than for Whites, although generally quite low for both groups.

Table 3: Using Race-Specific GLI Reference Equation

	Whites	Blacks	White - Black	p-value
Sensitivity				
Breathlessness	0.25	0.20	0.05	0.02
Dry Cough	0.18	0.17	0.01	0.92
Wheezing	0.27	0.21	0.06	0.02
Fair/Poor Health	0.25	0.16	0.10	0.00
Limited in Work	0.22	0.18	0.04	0.05
Specificity				
Breathlessness	0.92	0.91	0.01	0.36
Dry Cough	0.89	0.90	-0.01	0.27
Wheezing	0.91	0.91	0.00	0.81
Fair/Poor Health	0.91	0.91	0.00	0.89
Limited in Work	0.91	0.91	0.00	0.97
Positive Predictive Value				
Breathlessness	0.62	0.52	0.09	0.02
Dry Cough	0.08	0.08	0.01	0.85
Wheezing	0.36	0.29	0.07	0.02
Fair/Poor Health	0.31	0.35	-0.05	0.18
Limited in Work	0.36	0.31	0.05	0.12
Negative Predictive Value				
Breathlessness	0.71	0.70	0.01	0.62
Dry Cough	0.95	0.95	0.00	0.39
Wheezing	0.87	0.87	0.00	0.98
Fair/Poor Health	0.88	0.77	0.11	0.00
Limited in Work	0.83	0.83	0.01	0.57

The results in Table 4 indicate that using the White standard, or the GLI-2012 equation fit only to the healthy White population, does not increase fairness between Whites and Blacks. While sensitivity increases significantly for Blacks, specificity and positive predictive value both decrease. This is likely because the standards for Whites are lower than for Blacks, so using the White standard for Black people increases the chance of predicting someone to be unhealthy, regardless of their actual lung function. As such, it cannot be conclusively determined that the White standard is better, either in efficacy or in fairness.

Table 4: Using White GLI Reference Equation for Both Groups

	Whites	Blacks	White - Black	p-value
Sensitivity				
Breathlessness	0.25	0.49	-0.24	0.00
Dry Cough	0.18	0.44	-0.26	0.00
Wheezing	0.27	0.55	-0.29	0.00
Fair/Poor Health	0.25	0.44	-0.19	0.00
Limited in Work	0.22	0.47	-0.25	0.00
Specificity				
Breathlessness	0.92	0.66	0.26	0.00
Dry Cough	0.89	0.63	0.26	0.00
Wheezing	0.91	0.65	0.26	0.00
Fair/Poor Health	0.91	0.65	0.26	0.00
Limited in Work	0.91	0.64	0.27	0.00
Positive Predictive Value				
Breathlessness	0.62	0.41	0.20	0.00
Dry Cough	0.08	0.06	0.03	0.04
Wheezing	0.36	0.21	0.15	0.00
Fair/Poor Health	0.31	0.28	0.02	0.33
Limited in Work	0.36	0.23	0.13	0.00
Negative Predictive Value				
Breathlessness	0.71	0.73	-0.02	0.20
Dry Cough	0.95	0.96	-0.01	0.30
Wheezing	0.87	0.90	-0.02	0.01
Fair/Poor Health	0.88	0.79	0.10	0.00
Limited in Work	0.83	0.84	-0.01	0.47

Finally, Table 5 depicts the results of our third specification, which is the GLI-2012 equation (approximated by a 5th percentage quantile regression) trained on both Whites and Blacks. Even without the inclusion of race during training, we see that the model still remains unfair. In fact, the difference between groups is significant for nearly all combinations of outcomes and metrics. Note however that the efficacy of this model for Blacks is similar to that of the original GLI equation which factors in race for most metrics. As such, we can see that not including race does not improve or worsen outcomes for Blacks significantly across the board, but does however increase the gap between efficacy for both groups, worsening mathematical fairness.

Discussion

In this study, we aimed to understand the impact of race correction in pulmonary function tests. We found that two types of race neutral approaches, using White reference values and ignoring race altogether, do not lead to increased fairness for Blacks. Notably, we do find significant differences in fairness for sensitivity across equations.

Some recent research has found that race-specific reference values do not improve prediction for Blacks. Ekstrom and Mannino found that race-specific GLI-2012 equations do not improve prediction of breathlessness or mortality for Blacks when compared to using White-reference values[8]. Similarly, Elmaleh-Sachs et al. find that race-specific spirometry reference equations do not improve the

Table 5: Race Neutral Quantile Regression

	Whites	Blacks	White - Black	p-value
Sensitivity				
Breathlessness	0.12	0.24	0.19	0.00
Dry Cough	0.12	0.09	-0.75	0.31
Wheezing	0.10	0.30	0.19	0.00
Fair/Poor Health	0.13	0.20	0.16	0.00
Limited in Work	0.09	0.20	0.13	0.00
Specificity				
Breathlessness	0.98	0.87	0.55	0.00
Dry Cough	0.97	0.86	0.63	0.00
Wheezing	0.97	0.90	0.10	0.00
Fair/Poor Health	0.98	0.88	0.02	0.00
Limited in Work	0.97	0.88	-0.06	0.00
Positive Predictive Value				
Breathlessness	0.73	0.51	0.50	0.00
Dry Cough	0.15	0.04	-0.77	0.00
Wheezing	0.41	0.36	0.26	0.19
Fair/Poor Health	0.44	0.31	0.17	0.00
Limited in Work	0.41	0.30	0.19	0.01
Negative Predictive Value				
Breathlessness	0.68	0.68	0.08	0.84
Dry Cough	0.96	0.94	0.56	0.01
Wheezing	0.86	0.87	-0.11	0.11
Fair/Poor Health	0.88	0.80	0.77	0.00
Limited in Work	0.83	0.81	0.25	0.01

prediction of chronic lower respiratory disease events and mortality compared to a race neutral approach. Their race neutral appraoch is represented as an average over the race-specific reference values for Whites, Blacks, North East Asians, and South East Asians [7]. Lastly, a secondary data analysis of the CARDIA Lung Study found that a disproportionate number of Black participants had lung function readings that, when corrected for race, appeared "normal", but actually had apparent emphysema on their CT scans [10].

To our knowledge, our study is the first to investigate the use of a quantile regression fit as a proxy for a truly race-neutral reference equation as opposed to using White-reference values or an average over race-specific reference equations. Additional strengths of this study include the large sample and variety of outcomes investigated including measures of physical functional impairments.

There are several limitations associated with this study. First, we assess fairness using respondents to the NHANES 2007-2012 survey. Respondents may not be representative of the public as a whole and our findings may not be generalizable to the broader US or global population. In fact, we find that 11.21% of respondents have a low FEV1 value based on the GLI-2012 race-specific reference equation. In a healthy non-smoker population, one would expect this value to be closer to 5%. Second, pulmonary and physical impairments like breathlessness and fair/poor health may not be the best proxies for low levels of lung function. Third, we did not have direct measures of these variables and instead must rely on self-report. Fourth, we only assess fairness for Black and White participants. Further work is needed to understand how race-adjustment may lead to increased or decreased fairness for other minority groups including Native Americans, Hispanics/Latinos, and South, South-East, and East Asians.

The mechanisms by which race improves the prediction of lung function in these algorithms are not fully understood. It is possible that race is a proxy for other factors that are associated with lung function, such as genetics or environmental exposures. Further research is needed to fully understand the relationship between race and lung function, and how this relationship can be leveraged to improve healthcare outcomes.

Conclusion

In conclusion, the use of race in pulmonary function algorithms leads to increased fairness in the prediction of lung function. This finding has important implications for reducing racial disparities in lung function and improving healthcare outcomes for individuals of all racial backgrounds. Importantly, we find that there are clear gaps for improvement with respect to the sensitivity of these diagnostic tests. Further research is needed to evaluate mechanisms by which race improves the accuracy of lung function measurement, as well as to explore alternative approaches to addressing potential differences in test performance between racial groups.

Acknowledgments. We'd like to acknowledge the Harvard CS288 AI For Social Good course staff - Professor Milind Tambe, Sonja Johnson-Yu, and Paula Rodriguez Diaz - for their support, advice, and guidance.

Code and Data availability

All data used are publicly available from NHANES. Code used for all analyses is available here.

References

[1] "Sarah Beaverson BS, Victoria M. Ngo BS, Meera Pahuja MD, Alan Dow MD MSc, Patrick Nana-Sinkam MD MSHA, and Matthew Schefft DO MSHA". "race adjustments in calculating lung function from spirometry measurements". *Journal of Hospital Medicine*, October, 2022.

[2] "Nada Hassanein". "lung disease tests are failing black patients, studies show. experts are calling for change.". *USA*, October 12, 2022.

[3] Lundy Braun. Race, ethnicity and lung function: a brief history. *Canadian journal of respiratory therapy: CJRT= Revue canadienne de la therapie respiratoire: RCTR*, 51(4):99, 2015.

[4] Lucia Trimber. The nfl's reversal on 'race norming' reveals how pervasive medical racism remains. *NBC THINK*, 2021.

[5] "Carlos A. Vaz Fragoso, John Concato, Gail McAvay, Peter H. Van Ness, Carolyn L. Rochester, H. Klar Yaggi, and Thomas M. Gill". "the ratio of fev1 to fvc as a basis for establishing chronic obstructive pulmonary disease". *American Journal of Respiratory and Critical Care Medicine*, 2010.

[6] Philip H. Quanjer, Sanja Stanojevic, Tim J. Cole, Xaver Baur, Graham L. Hall, Bruce H. Culver, Paul L. Enright, John L. Hankinson, Mary S.M. Ip, Jinping Zheng, and Janet Stocks. Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global

lung function 2012 equations. *"European Respiratory Journal"*, 40:1324 – 1343, 2012.

[7] Arielle Elmaleh-Sachs, Pallavi Balte, Elizabeth C. Oelsner, Norrina B. Allen, Aaron Baugh, Alain G. Bertoni, John L. Hankinson, Jim Pankow, Wendy S. Post, Joseph E. Schwartz, and et al. Race/ethnicity, spirometry reference equations, and prediction of incident clinical events: The multi-ethnic study of atherosclerosis (mesa) lung study. *American Journal of Respiratory and Critical Care Medicine*, 205(6):700–710, 2022.

[8] Magnus Ekström and David Mannino. Research race-specific reference values and lung function impairment, breathlessness and prognosis: Analysis of nhanes 2007–2012. *Respiratory research*, 23(1):1–8, 2022.

[9] Roger Koenker and Gilbert Bassett. Regression quantiles. *Econometrica*, 46(1):33, 1978.

[10] Gabrielle Liu, Sadiya Khan, Laura Colangelo, Daniel Meza, George Washko, Peter Sporn, David Jacobs Jr., Mark Dransfield, Mercedes Carnethon, and Ravi Kalhan. Comparing racial differences in emphysema prevalence among adults with normal spirometry: A secondary data analysis of the cardia lung study. *Annals of Internal Medicine*, 2022.