Published in final edited form as:

Annu Rev Med. 2023 January 27; 74: 401–412. doi:10.1146/annurev-med-043021-024004.

Origins of Racial and Ethnic Bias in Pulmonary Technologies

Michael W. Sjoding[1,2,3], **Sardar Ansari**[2,4], **Thomas S. Valley**[1,3,5]

[1]Division of Pulmonary and Critical Care Medicine, Department of Internal Medicine, University of Michigan Medical School, Ann Arbor, Michigan, USA

[2]Weil Institute for Critical Care Research and Innovation, Ann Arbor, Michigan, USA

[3]Institute for Healthcare Policy and Innovation, University of Michigan, Ann Arbor, Michigan, USA

[4]Department of Emergency Medicine, University of Michigan Medical School, Ann Arbor, Michigan, USA

[5]Center for Bioethics and Social Sciences in Medicine, University of Michigan, Ann Arbor, Michigan, USA

Abstract

Understanding how biases originate in medical technologies and developing safeguards to identify, mitigate, and remove their harms are essential to ensuring equal performance in all individuals. Drawing upon examples from pulmonary medicine, this article describes how bias can be introduced in the physical aspects of the technology design, via unrepresentative data, or by conflation of biological with social determinants of health. It then can be perpetuated by inadequate evaluation and regulatory standards. Research demonstrates that pulse oximeters perform differently depending on patient race and ethnicity. Pulmonary function testing and algorithms used to predict healthcare needs are two additional examples of medical technologies with racial and ethnic biases that may perpetuate health disparities.

Keywords

disparities; pulse oximetry; structural racism; pulmonary function testing; medical devices; algorithms

INTRODUCTION

Naïve, conventional wisdom suggests that machines cannot be biased because they are objective, inanimate objects that lack the ability to make conscious decisions (1, 2). Yet, limitations in the design of hardware or software can result in systematic performance differences in populations based on attributes such as race, ethnicity, gender, sex, or socioeconomic status. For example, biased hardware designs within automated soap

msjoding@umich.edu .

DISCLOSURE STATEMENT
The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

dispensers result in technology that readily dispenses soap to individuals with light skin tones but fails to dispense soap to individuals with darker skin tones (3). High-profile examples of software technologies that may perpetuate discriminatory practices include algorithms used for facial recognition (4), loan decisions (5), and criminal sentencing (6). This recognition of racial bias within technologies incorporates the modern understanding of racial bias as rooted not only in explicit individual prejudices or racism, but also in systems, laws, policies, or practices in the form of structural racism, whether they are intentionally biased or not (7, 8).

Health care is not immune to these critical problems. Bias is well documented in medical practice, affecting behavior, interactions, and decision making, where it may play a role in perpetuating health disparities (9, 10). Analogously, medical devices can also exhibit racial or ethnic bias if design flaws lead to performance differences in patients of racial or ethnic minority groups (2). While these design flaws may largely be unintentional, it is incumbent upon designers and users to make every effort to identify, mitigate, and remove these biases so that they do not contribute to the stark health disparities of minority groups (11, 12).

Pulse oximeter technology serves as an important case study in how bias can be introduced, be perpetuated, and remain unaddressed in medical devices. In this article, we examine the historical development of the pulse oximeter and describe recent research highlighting performance differences by patient race and ethnicity. We then compare pulse oximeters with two other medical technologies recently recognized as perpetuating health disparities: pulmonary function testing and algorithms used to predict healthcare need. We identify key points where bias can be introduced, perpetuated, and addressed in medical technologies and discuss general strategies to ensure that medical technologies perform equivalently on all patients.

THE PULSE OXIMETER: BIAS INTRODUCED IN THE PHYSICAL DESIGN OF MEDICAL TECHNOLOGY

The modern pulse oximeter noninvasively measures the oxygen content of arterial blood. This medical device has been described as “arguably the most significant technological advance ever made in monitoring the well-being and safety of patients during anesthesia, recovery, and critical care” (13, p. 285). Blood oxygen level is considered as important as other cardinal vital signs such as temperature or blood pressure (14). Before the development of modern noninvasive pulse oximeters, measurement of blood oxygen levels was time-consuming, painful, and unreliable. The modern pulse oximeter was first conceived in 1972 by Takuo Aoyagi, who recognized that by measuring the pulsatile change of red and infrared (IR) light absorbance through tissue, oxygen saturation could be computed (15). By the 1980s, pulse oximeters were widely adopted as the standard of care for patient monitoring in hospitals and clinics. At the start of the COVID-19 pandemic, pulse oximeters rose to further prominence as an essential home monitoring device for patients with COVID-19 (16).

The common finger pulse oximeter works by shining light through the fingertip at two wavelengths, approximately 660 nm (red) and 940 nm (IR), and measuring the light transmitted across the finger. These two wavelengths are chosen because oxygenated blood and deoxygenated blood absorb light to a different degree at these wavelengths. The Beer–Lambert law describes how light transmission through a dissolved substance is related to the concentration of the substance. The pulse oximeter capitalizes on this principle by measuring the amount of light transmitted through the finger at both light wavelengths to determine the concentration of oxygenated blood.

When a pulse oximeter sends light through the finger, most light is absorbed or scattered by tissue, bone, and venous blood. Each heartbeat causes a small increase in arterial blood into the finger, increasing the level of light absorbed at both wavelengths. Because the device isolates the arterial change in light absorption due to pulsating blood, it was generally understood that pulse oximeter accuracy was not influenced to a meaningful degree by static factors of the finger such as skin thickness, subcutaneous fat, or skin tone (17).

As the pulse oximeter continuously measures light transmission at both wavelengths, signal processing algorithms in the device identify the peak and trough light transmittance during each cardiac cycle to isolate the absorbance due to arterial blood. The AC component of the photoplethysmography signal is associated with the pulsating arterial blood while the DC component is associated with light absorption by tissue, venous blood, and nonpulsating arterial blood. The peak and trough amplitudes of the photoplethysmography signal are used to calculate the AC and DC components at both wavelengths and the modulation ratio: $R = (AC_{red}/DC_{red})/(AC_{IR}/DC_{IR})$ (18).

There is no direct mathematical relationship between R and the arterial oxygen saturation of blood (SaO_2). Therefore, pulse oximeter manufacturers empirically determine the relationship between R and SaO_2 for a device on the basis of data collected from test subjects. This is commonly done by measuring R in healthy volunteers whose saturations were altered from 100% to approximately 70% by breathing various hypoxic gas mixtures (17). An underlying assumption of this approach is that the relationship between the measured parameter R and physiological parameter SaO_2 did not have clinically meaningful variability between test subjects and the general population.

In 1990, Jubran & Tobin (19) published one of the first studies to describe differences in pulse oximeter accuracy based on race. They studied whether pulse oximeters could be used to safely titrate oxygen in patients receiving invasive mechanical ventilation. They compared oxygen saturation measured by pulse oximeters (SpO_2) with arterial oxygen tension (PaO_2) and SaO_2 directly measured from simultaneously collected arterial blood gas samples in a cohort of 25 White and 29 Black critically ill patients. They found that pulse oximeters overestimated oxygen saturation to a greater degree in Black patients than in White patients. Pulse oximeter bias, calculated as the average difference between SpO_2 and SaO_2 readings, was significantly higher in Black patients (3.3 ± 2.7%) than in White patients (2.2 ± 1.8%). The authors concluded that the optimal SpO_2 target to ensure a safe level of oxygen in the blood ($PaO_2 \geq 60$) should differ based on a patient's race, with a minimum SpO_2 of 95% in Black patients and a minimum SpO_2 of 92% in White patients.

Two laboratory-based investigations of healthy volunteers conducted in 2005 and 2007 also found that pulse oximeters overestimated arterial oxygen levels in individuals with darker skin tone. The 2005 study compared pulse oximeter values in 10 White and 11 Black patients, finding that skin pigment–related bias increased approximately in proportion to the level of desaturation (20). The 2007 study further explored the issue by grouping 36 subjects with a range of skin tones into “light,” “intermediate,” and “dark” skin pigmentation groups, identifying a dose-response between the degree of skin pigment and pulse oximeter bias (21). However, both studies concluded that the magnitude of pulse oximeter error related to skin pigmentation was relatively small at SaO_2 values above 80% and probably of no general clinical significance.

A 2020 study analyzed retrospective electronic health record data of SpO_2 and SaO_2 samples collected during routine practice from patients hospitalized in a single center in 2020 and from 178 US hospitals in 2014 and 2015 (22). Analyzing routinely collected data provided the opportunity to analyze samples from 1,333 White and 276 Black patients in the single-center cohort and 7,342 White and 1,050 Black patients in the multi-center cohort—an order of magnitude more measurements than prior studies. SpO_2 measurements were compared to subsequent arterial blood samples if they were collected within 10 min. In the primary analysis, the study investigated how frequently pulse oximeter measurements between 92% and 96% missed true oxygen levels below 88%, which they termed “occult hypoxemia.” A SpO_2 range of 92–96% was chosen because changes in care would be less likely to occur in the time between SpO_2 and SaO_2 measurements for oxygen saturation measurements in that range. SpO_2 measurements higher than 96% may result in clinical changes to reduce oxygen supplementation, whereas SpO_2 measurements below 92% may result in clinical changes to increase oxygen supplementation. The study found that occult hypoxemia was three times more common in measurements from Black patients than in measurements from White patients.

The primary analysis performed in the 2020 study was a departure from typical metrics used to evaluate pulse oximeter accuracy, such as bias and accuracy root mean square error (Arms) (23). Pulse oximeter bias is the mean difference between SpO_2 and SaO_2 values, and Arms is the square root of the mean of squared differences between SpO_2 and SaO_2 values (23). The 2020 study measured how often low SaO_2 values occurred despite normal SpO_2 values, an analysis that was more similar to the Jubran & Tobin (19) study, which tried to identify safe pulse oximeter saturation targets that minimized hypoxemia. This framing may be better aligned with how pulse oximeters are used in clinical practice. Clinicians primarily make decisions based on pulse oximeter readings as a surrogate for, and without knowledge of, true SaO_2. The metric is easy to comprehend and represents how often pulse oximeters with a normal reading miss low oxygen levels. In diagnostic testing, this metric is called the false omission rate and is equal to the number of false negatives divided by the total number of true negatives and false negatives.

Compared to laboratory-based studies evaluating pulse oximeter accuracy, such as the two published in 2005 and 2007, the 2020 study had several important limitations. Because SpO_2 and SaO_2 values were not collected synchronously, some fluctuation in the arterial saturation would be expected within the time frame between a SpO_2 measurement and arterial blood gas collection. This could result in the appearance of lower pulse oximeter accuracy compared to laboratory studies. However, this would not explain the differences seen between Black and White patients unless clinicians reacted in a systematically different fashion for Black and White patients in the time between SpO_2 and SaO_2 measurements. The 2020 study also relied on self-reported race, rather than an objective measurement of skin tone. Race is a social construct used to categorize populations primarily based on physical traits or ancestry (24). While differences in skin tone are present across racial groups, variation is also present within groups (25). For example, in the National Survey of Black Americans, 1979–1980, professionally trained Black interviewers categorized skin complexion as “very dark” in 8.5% of respondents and “light or very light” in 17% of respondents (26). Despite these limitations, the 2020 study raised significant concerns about bias in pulse oximeters, leading the US Food and Drug Administration (FDA) to issue a communication in February 2021 warning about the use of such devices for clinical decision making (27).

After the 2020 study, several others demonstrated differences in pulse oximeter accuracy between Black and White patients. An analysis of critically ill patients undergoing evaluation for extracorporeal membrane oxygenation found that self-reported Black patients had twice the rate of occult hypoxemia compared to White patients but did not find differences between White, Hispanic, and Asian patients (28). Another study of 87,971 patient encounters found that pulse oximeters missed occult hypoxemia most often in Black, followed by Hispanic, Asian, and White patients (29). This same study found that patients with occult hypoxemia subsequently had more organ dysfunction and higher in-hospital mortality, despite clinically similar appearance at hospital presentation. The association between occult hypoxemia and mortality was replicated in a separate study of patients admitted to the intensive care unit (ICU) or undergoing surgery (30). These latter two studies highlight how differences in pulse oximeter accuracy across racial groups could contribute to disparities in patient outcomes.

PULMONARY FUNCTION TESTING: BIAS IN THE INTERPRETATION OF MEASUREMENTS FROM MEDICAL TECHNOLOGY

Pulmonary function testing is a common procedure used to measure how much and how quickly air can be moved in and out of the lungs during inspiration and expiration. This series of tests is commonly used to diagnose and quantify the severity of lung diseases (31). Racial and ethnic disparities in pulmonary function testing, in contrast to pulse oximetry, are not due to the inherent design of the device used to measure lung function. Instead, disparities have arisen because of how the results are interpreted.

Interpreting pulmonary function test results and determining whether values are normal is challenging because important characteristics such as age, sex, and height can influence the size and elasticity of normal, healthy lungs. Therefore, an individual patient’s test results are compared to average values taken from population-based studies to determine whether they are normal (32). Reference equations for normal pulmonary function values have been derived using large population-based studies of nonsmoking individuals.

In large population-based studies, Black and Hispanic patients of the same age, sex, and height typically have lower average lung function than White patients (33, 34). Rather than considering how potential differences in social and environmental exposures might affect lung development and lead to these differences, lung function differences were often interpreted in the context of scientifically inaccurate beliefs that Black individuals had smaller lungs than White individuals. Such beliefs were described in writings from Thomas Jefferson in the 1700s and reinforced by faulty scientific methods during the 1800s (35). Race- and ethnicity-based lung function equations or correction factors that assume a 10–15% lower lung capacity for Black patients have been common practice in the use of modern pulmonary function tests since the 1970s.

Mirroring broader conversations about race-based correction in clinical algorithms (36), concerns have been raised about race-based correction to pulmonary function testing and its potential to exacerbate disparities (37). In a 2022 study, Baugh et al. (38) evaluated how well lung function equations with and without race-based correction correlated with other measures of lung health. They found that equations without race-based correction better aligned with respiratory symptoms, as quantified by the St. George’s Respiratory Questionnaire (39), and with airway wall thickness (40), a measure of airway disease. They concluded that the use of race-based correction may normalize lung injury due to long-standing discrimination against minorities. Another recent study concluded that lung function equations with race-based correction did not improve the prediction of respiratory disease events or mortality compared to equations without race-based correction, further questioning the clinical utility of such equations (41).

The use of race-based correction in pulmonary function testing is problematic for several reasons. First, overestimating the lung function of Black individuals may lead to underdiagnosis and undertreatment of lung disease. Race-based correction artificially boosts lung function measurements for Black patients up to 15%. As a result, Black individuals with respiratory symptoms may be falsely reassured by seemingly normal lung function values. Second, race-based correction perpetuates centuries of racist beliefs that Black individuals were inferior and lacked fitness (42). Finally, race-based correction presumes race and ethnicity to be a biological, rather than a social, construct, bringing inaccuracy to the interpretation of pulmonary function testing. By inappropriately conflating social and environmental factors that influence lung development with inherent biological differences between races and ethnicities, the use of race-based correction distorts accurate measurements from a pulmonary function device.

ALGORITHMS USED TO PREDICT HEALTHCARE NEED: BIAS IN DATA USED TO TRAIN MEDICAL SOFTWARE

In contrast to pulse oximeters and pulmonary function testing, algorithms used to manage the health of populations and determine healthcare needs are solely software-based medical technologies. Large health systems, insurance companies, and governmental agencies have widely adopted commercial risk-prediction algorithms (43). One use of these algorithms is to identify high-risk patients who may benefit from complex care management programs, with the goal of providing additional resources to these medically complex patients before their health further deteriorates. Because these programs themselves are costly to operate, health systems may use commercial algorithms to help select patients for these programs (44).

A 2019 study by Obermeyer et al. (45) demonstrated how a widely used commercial risk-prediction algorithm resulted in racial bias. The algorithm was designed to estimate future healthcare needs using previous healthcare insurance claims data, insurance type, diagnosis and procedure codes, prescribed medications, and detailed healthcare encounter billed amounts, but it did not include race as a predictor variable. The researchers compared predictions made by the algorithm to other markers of patient health, including the number of active chronic medical conditions, blood pressure control, diabetes severity measured by hemoglobin A1c, kidney function measured by creatinine, low-density lipoprotein cholesterol, and anemia. Black patients classified in the same risk category as White patients were found to be sicker by all measures. This bias in classification resulted in fewer Black patients being referred for complex care management programs.

The primary reason for the algorithm’s racial bias was that its design conflated healthcare costs with healthcare needs (45, 46). Unequal access to care results in less healthcare spending for Black patients at the same level of illness as White patients. Thus, the algorithm learned the historical inequalities present in the data used for training, falsely concluding that Black patients were healthier than White patients because fewer healthcare dollars were spent on them. When researchers studied a new algorithm trained to predict a patient’s number of active chronic medical conditions rather than their healthcare costs, the fraction of Black patients above the threshold for automatic referral to the complex care management program nearly doubled, increasing from 14% to 27%.

While reducing the bias in the clinical risk prediction software may be as simple as rerunning the algorithm with another outcome label, ensuring that similar problems do not arise in other clinical risk algorithms is decidedly more difficult. When developing a software algorithm, Obermeyer et al. (45) highlighted the essential step of careful problem formulation, which is the translation of high-level objectives or strategic goals into a tractable problem that can be solved with available data (47). Predicted cost is an easily measurable outcome, common among other risk algorithms, and was even suggested in the literature as a method for identifying high-need patients (48). Yet, a lack of nuanced understanding of the inherent bias in using healthcare costs as an outcome variable led to the development of a racially biased algorithm. Because software algorithms are typically trained using historical data that are often biased by human decisions and systemic racism, it is critical that algorithm designers have a deep appreciation for the social and historical influences on these data (49).

HOW BIAS IS INTRODUCED AND PERPETUATEDIN MEDICAL TECHNOLOGIES

Bias impacting racial or ethnic groups can be introduced and perpetuated at several points during medical technology development (Table 1). Frequently, failures occur at multiple stages. Establishing safeguards at each stage could work toward limiting bias in new medical technologies.

Physical Design of the Technology

Fundamental aspects of the pulse oximeter's physical design contribute to differences in performance across patient groups. Important individual characteristics such as skin pigment, finger size, fingernail polish, and skin perfusion are relevant to pulse oximeter accuracy (50). Because pulse oximeters were designed to isolate light absorption from pulsatile blood, it was assumed that pulse oximeter accuracy would not be significantly influenced by static factors of the finger such as skin tone. Yet, this assumption has proved to be inaccurate. Finger size, which impacts light absorption and scatter, may also impact accuracy. Difference in finger size is a hypothesized reason for accuracy differences between males and females or adults and children (21, 51, 52).

Concerns have also been raised about the physical design of other medical devices such as orthopedic implants and cardiac pacemakers (53–55). If such technologies were primarily designed for the anatomy of an average male, this could result in a higher failure rate in females and in males of small stature. Similarly, if pulse oximeters were originally designed and tested on patients with almost universally lighter skin tone, inaccuracies in subjects of more diverse backgrounds may go underappreciated. Knowledge of whether a particular device design may introduce bias among key patient subgroups may be difficult to obtain in the early design phase. Ensuring diversity in the engineering design team may be one possible mitigation strategy.

Use of Data That Do Not Reflect Important Subgroups

Biased data used to develop medical technologies are a common root cause of performance variation across racial and ethnic groups. A high-profile example of underrepresentative data as a source of bias was identified during the evaluation of several facial recognition technologies (4). One of the most commonly used face databases, Labeled Faces in the Wild, was estimated to be 78% male and 84% White (56). Using such a database to develop facial recognition technology resulted in significantly lower software performance in Black females. A similar issue would arise in pulse oximeters if the original calibration of the modulation ratio to SaO_2 was performed in a homogeneous population of healthy, light-skinned males. Given the importance of skin pigment and finger size to pulse oximeter accuracy, ensuring that the initial development population has significant diversity in these characteristics is critical. Including an equal distribution of males and females of both light and dark pigmentation might ensure that pulse oximeters perform more equally in these groups.

Conflation of Biological and Social Determinants of Health

Data used to establish normal values for pulmonary function testing and algorithms to allocate complex care management illustrate the need to critically evaluate why underlying racial or ethnic differences in the data exist. These differences, when identified at the population level, are increasingly recognized as reflective of the lived experiences of minority groups rather than reflective of true biological differences (57). Definitions of race and ethnicity, both clinically and in medical research, often rely on characteristics unrelated to biology and genetics, such as physical appearance, language, culture, and religion (58). The extent of meaningful genetic differences is highly questionable, with studies suggesting more variation within racial groups than between groups (59).

Race, ethnicity, biological determinants of health (e.g., genetics), and social and environmental determinants of health (e.g., poverty, medical access, blight, pollution) are often conflated. The use of race-based correction factors (as in pulmonary or kidney function testing) arose from racial or ethnic differences identified at the population level that were inappropriately attributed to biological rather than social and environmental origins (36). Similarly, complex care management algorithms that used healthcare costs as a proxy for healthcare needs failed to consider the social determinants (such as medical access) that contribute to healthcare costs and disparities across racial groups. However, there have been several recent efforts to acknowledge and eliminate these sources of racial and ethnic disparities. Work supported by the National Kidney Foundation and the American Society of Nephrology led to new clinical algorithms for estimating kidney function that removed race-based correction (60). Researchers affiliated with the Maternal-Fetal Medicine Units Network also recently updated a risk tool predicting the success rate of vaginal birth after a cesarean delivery by removing variables for race and ethnicity (61).

Imperfect Technology Evaluation and Regulatory Standards

Medical device regulation and rigorous peer-reviewed medical literature are essential safeguards to mitigate the impact of biased medical technologies. While the United States and Europe well-established systems for medical device regulation, concerns have been raised about their effectiveness (62). A review of the highest risk medical device applications submitted to the FDA in 2014–2017 found that device performance reporting for safety and efficacy by gender, race, or age is uncommon (63). When subgroups are reported, the number of patients is often too small to enable meaningful conclusions. Current FDA 510(k) premarket notification guidance for pulse oximeters recommends that performance should be reported on 10 or more healthy subjects that vary in age and gender (64). The guidance states that at least 2 subjects or 15% of the subject pool, whichever is larger, should have darkly pigmented skin. The small number of darkly pigmented subjects and lack of specific subgroup analysis in this recommendation has been cited as a critical regulatory gap—insufficient to ensure that pulse oximeters perform equivalently on darkly and lightly pigmented patients (65).

Regulation of clinical algorithms or software used to support medical decisions is more variable. The International Medical Device Regulators Forum developed a regulatory framework to evaluate software as a medical device, basing the level of regulation on the level of risk to patients (66). For instance, software used to diagnose and treat patients in critical healthcare settings requires the highest level of scrutiny, while software used to inform management in nonserious settings receives substantially less scrutiny (67). Software algorithms used to estimate patient risk, such as the one evaluated by Obermeyer et al. (45), or equations used to determine normal pulmonary or kidney function are not strictly regulated. Thus, independent peer-reviewed medical research and the efforts of medical societies are the primary way these technologies are evaluated.

Recognizing that clinical algorithms can be biased, understanding the mechanisms of bias, and promoting efforts to evaluate technologies for bias are essential steps toward preventing racial and ethnic disparities. Grassroots efforts such as the Algorithmic Justice League worked to remove biased facial recognition software (https://www.ajl.org/learn-more). Independent initiatives to ensure that newly developed medical technologies are unbiased could serve the same purpose in health care.

CONCLUSION

Disparities in medical technology performance can be introduced and perpetuated at several points throughout its development. Greater interest and awareness in these issues echo broader conversations about the roots of structural racism in health care and society (68, 69). Understanding how biases in medical technologies originate and developing safeguards to identify, mitigate, and remove their harms are essential to ensuring that medical technologies perform equivalently for all individuals.

ACKNOWLEDGMENTS

The authors thank Theodore J. Iwashyna for his review and feedback on an earlier version of this manuscript.

LITERATURE CITED

1. Sjoding M, Iwashyna TJ, Valley TS. 2021. More on racial bias in pulse oximetry measurement. Reply. N. Engl. J. Med 384:1278
2. Benjamin R 2019. Race After Technology: Abolitionist Tools for the New Jim Code. Medford, MA: Polity
3. Fussell S 2017. Why can’t this soap dispenser identify dark skin? Gizmodo. https://gizmodo.com/why-cant-this-soap-dispenser-identify-dark-skin-1797931773
4. Buolamwini J, Gebru T. 2018. Gender shades: intersectional accuracy disparities in commercial gender classification. Paper presented at the Conference on Fairness, Accountability, and Transparency, New York, NY, Feb. 23–24
5. Bartlett R, Morse A, Stanton R, Wallace N. 2019. Consumer-lending discrimination in the FinTech era. NBER Work. Pap. 25943
6. Angwin J, Larson J, Mattu S, Kirchner L. 2016. Machine bias. Propublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
7. Braveman PA, Arkin E, Proctor D, et al. 2022. Systemic and structural racism: definitions, examples, health damages, and approaches to dismantling. Health Aff. 41:171–78
8. Bailey ZD, Feldman JM, Bassett MT. 2021. How structural racism works—racist policies as a root cause of U.S. racial health inequities. N. Engl. J. Med 384:768–73 [PubMed: 33326717]
9. Chapman EN, Kaatz A, Carnes M. 2013. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. J. Gen. Intern. Med 28:1504–10 [PubMed: 23576243]
10. Smedley BD, Stith AY, Nelson AR, eds. 2003. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: Natl. Acad. Press
11. Braveman P 2014. What are health disparities and health equity? We need to be clear. Public Health Rep. 129(Suppl. 2):5–8
12. AHRQ. 2019. 2018 National healthcare quality and disparities report. Agency Healthc. Res. Qual., Rockville, MD. https://www.ahrq.gov/research/findings/nhqrdr/nhqdr18/index.html

13. Severinghaus JW, Astrup PB. 1986. History of blood gas analysis. VI. Oximetry. J. Clin. Monitor 2:270–884

14. Neff TA. 1988. Routine oximetry. A fifth vital sign? Chest 94:227

15. Severinghaus JW. 2007. Takuo Aoyagi: discovery of pulse oximetry. Anesth. Analg 105:S1–S4 [PubMed: 18048890]

16. Greenhalgh T, Knight M, Inada-Kim M, et al. 2021. Remote management of covid-19 using home pulse oximetry and virtual ward support. BMJ 372:n677 [PubMed: 33766809]

17. Schnapp LM, Cohen NH. 1990. Pulse oximetry. Uses and abuses. Chest 98:1244–50 [PubMed: 2225973]

18. Mannheimer PD. 2007. The light-tissue interaction of pulse oximetry. Anesth. Analg 105:S10–S17 [PubMed: 18048891]

19. Jubran A, Tobin MJ. 1990. Reliability of pulse oximetry in titrating supplemental oxygen therapy in ventilator-dependent patients. Chest 97:1420–25 [PubMed: 2347228]

20. Bickler PE, Feiner JR, Severinghaus JW. 2005. Effects of skin pigmentation on pulse oximeter accuracy at low saturation. Anesthesiology 102:715–19 [PubMed: 15791098]

21. Feiner JR, Severinghaus JW, Bickler PE. 2007. Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: the effects of oximeter probe type and gender. Anesth. Analg 105:S18–S23 [PubMed: 18048893]

22. Sjoding MW, Dickson RP, Iwashyna TJ, et al. 2020. Racial bias in pulse oximetry measurement. N. Engl. J. Med 383:2477–78 [PubMed: 33326721]

23. Batchelder PB, Raley DM. 2007. Maximizing the laboratory setting for testing devices and understanding statistical output in pulse oximetry. Anesth. Analg 105:S85–S94 [PubMed: 18048904]

24. Natl. Hum. Genome Res. Inst. 2022. Race. Natl. Hum. Genom Res. Inst. Talking Glossary of Genetic Terms https://www.genome.gov/genetics-glossary/Race. Accessed Mar. 21, 2022

25. Taylor SC. 2002. Skin of color: biology, structure, function, and implications for dermatologic disease. J. Am. Acad. Dermatol 46:S41–S62 [PubMed: 11807469]

26. Keith VM, Herring C. 1991. Skin tone and stratification in the Black community. Am. J. Sociol 97:760–78

27. FDA. 2021. Pulse oximeter accuracy and limitations: FDA safety communication. Saf. Commun., US Food Drug Adm., Silver Spring, MD. https://www.fda.gov/medical-devices/safety-communications/pulse-oximeter-accuracy-and-limitations-fda-safety-communication

28. Valbuena VSM, Barbaro RP, Claar D, et al. 2021. Racial bias in pulse oximetry measurement among patients about to undergo extracorporeal membrane oxygenation in 2019–2020: a retrospective cohort study. Chest 161:971–78 [PubMed: 34592317]

29. Wong AI, Charpignon M, Kim H, et al. 2021. Analysis of discrepancies between pulse oximetry and arterial oxygen saturation measurements by race and ethnicity and association with organ dysfunction and mortality. JAMA 4:e2131674

30. Henry NR, Hanson AC, Schulte PJ, et al. 2022. Disparities in hypoxemia detection by pulse oximetry across self-identified racial groups and associations with clinical outcomes. Crit. Care Med 50:204–11 [PubMed: 35100193]

31. Johnson JD, Theurer WM. 2014. A stepwise approach to the interpretation of pulmonary function tests. Am. Fam. Phys 89:359–66

32. Pellegrino R, Viegi G, Brusasco V, et al. 2005. Interpretative strategies for lung function tests. Eur. Respir. J 26:948–68 [PubMed: 16264058]

33. Hankinson JL, Odencrantz JR, Fedan KB. 1999. Spirometric reference values from a sample of the general U.S. population. Am. J. Respir. Crit. Care Med 159:179–87 [PubMed: 9872837]

34. Quanjer PH, Stanojevic S, Cole TJ, et al. 2012. Multi-ethnic reference values for spirometry for the 3–95-yr age range: the global lung function 2012 equations. Eur. Respir. J 40:1324–43 [PubMed: 22743675]

35. Braun L 2015. Race, ethnicity and lung function: a brief history. Can. J. Respir. Ther 51:99–101 [PubMed: 26566381]

36. Vyas DA, Eisenstein LG, Jones DS. 2020. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. N. Engl. J. Med 383:874–82 [PubMed: 32853499]

37. Schluger NW, Dozor AJ, Jung YEG. 2022. Rethinking the race adjustment in pulmonary function testing. Ann. Am. Thorac. Soc 19:353–56 [PubMed: 34784493]

38. Baugh AD, Shiboski S, Hansel NN, et al. 2021. Reconsidering the utility of race-specific lung function prediction equations. Am. J. Respir. Crit. Care Med 205:819–29

39. Am. Thorac. Soc. St. George’s Respiratory Questionnaire. American Thoracic Society Sleep Related Questionnaires. https://www.thoracic.org/members/assemblies/assemblies/srn/questionaires/sgrq.php

40. Patel BD, Coxson HO, Pillai SG, et al. 2008. Airway wall thickening and emphysema show independent familial aggregation in chronic obstructive pulmonary disease. Am. J. Respir. Crit. Care Med 178:500–5 [PubMed: 18565956]

41. Elmaleh-Sachs A, Balte P, Oelsener EC, et al. 2022. Race/ethnicity, spirometry reference equations, and prediction of incident clinical events: the Multi-Ethnic Study of Atherosclerosis (MESA) lung study. Am. J. Respir. Crit. Care Med 205:700–10 [PubMed: 34913853]

42. Braun L 2021. Race correction and spirometry: why history matters. Chest 159:1670–75 [PubMed: 33263290]

43. Hileman G, Steele S. 2016. Accuracy of claims-based risk scoring models: Society of Actuaries. https://www.soa.org/globalassets/assets/Files/Research/research-2016-accuracy-claims-based-risk-scoring-models.pdf

44. Bates DW, Saria S, Ohno-Machado L, et al. 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff. 33:1123–31

45. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366:447–53 [PubMed: 31649194]

46. Wiens J, Price WN 2nd, Sjoding MW. 2020. Diagnosing bias in data-driven algorithms for healthcare. Nat. Med 26:25–26 [PubMed: 31932798]

47. Passi S, Barocas S. 2019. Problem formulation and fairness. Paper presented at the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, Jan. 29–31

48. Long P, Abrams M, Milstein A, et al. 2017. Effective Care for High-Need Patients: Opportunities for Improving Outcomes, Value, and Health. Washington, DC: Natl. Acad. Med.

49. Benjamin R 2019. Assessing risk, automating racism. Science 366:421–22 [PubMed: 31649182]

50. Jubran A 2015. Pulse oximetry. Crit. Care 19:272 [PubMed: 26179876]

51. Ross PA, Newth CJ, Khemani RG. 2014. Accuracy of pulse oximetry in children. Pediatrics 133:22–29 [PubMed: 24344108]

52. Andrist E, Nuppnau M, Barbaro RP, et al. 2022. Association of race with pulse oximetry accuracy in hospitalized children. JAMA 5:e224584

53. Inacio MC, Ake CF, Paxton EW, et al. 2013. Sex and risk of hip implant failure: assessing total hip arthroplasty outcomes in the United States. JAMA Intern. Med 173:435–41 [PubMed: 23420484]

54. Chen A, Paxton L, Zhen X, et al. 2021. Association of sex with risk of 2-year revision among patients undergoing total hip arthroplasty. JAMA 4:e2110687

55. Musher DM, Rueda AM, Kaka AS, Mapara SM. 2007. The association between pneumococcal pneumonia and acute cardiac events. Clin. Infect. Dis 45:158–65 [PubMed: 17578773]

56. Han H, Jain AK. 2014. Age, gender and race estimation from unconstrained face images. Tech. Rep., Dep. Comput. Sci. Eng., Mich. State Univ., East Lansing, MI. http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation_MSUTechReport2014.pdf

57. Graves JL. 2018. Biological theories of race beyond the millennium. In Reconsidering Race: Social Science Perspectives on Racial Categories in the Age of Genomics, ed. Suzuki K, Von Vacano D, pp. 21–31. New York: Oxford Univ. Press

58. Deyrup A, Graves JL Jr. 2022. Racial biology and medical misconceptions. N. Engl. J. Med 386:501–3 [PubMed: 35119803]

59. Am. Assoc. Biol. Anthropol. 2019. AABA statement on race and racism. Position statement, Am. Assoc. Biol. Anthropol., Herndon, VA. https://physanth.org/about/position-statements/aapa-statement-race-and-racism-2019/

60. Inker LA, Eneanya ND, Coresh J, et al. 2021. New creatinine- and cystatin c-based equations to estimate GFR without race. N. Engl. J. Med 385:1737–49 [PubMed: 34554658]

61. Grobman WA, Sandoval G, Murguia Rice M, et al. 2021. Prediction of vaginal birth after cesarean delivery in term gestations: a calculator without race and ethnicity. Am. J. Obstet. Gynecol 225:664.e661–e667

62. Sorenson C, Drummond M. 2014. Improving medical device regulation: the United States and Europe in perspective. Milbank Q. 92:114–50 [PubMed: 24597558]

63. Fox-Rawlings SR, Gottschalk LB, Doamekpor LA, Zuckerman DM. 2018. Diversity in medical device clinical trials: Do we know what works for which patients? Milbank Q. 96:499–529 [PubMed: 30203600]

64. FDA. Pulse oximeters—premarket notification submissions [510(k)s]: guidance for industry and Food and Drug Administration staff. Guidance, US Food Drug Adm., Silver Spring, MD. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/pulse-oximeters-premarketnotification-submissions-510ks-guidance-industry-and-food-and-drug

65. Okunlola OE, Lipnick MS, Batchelder PB, et al. 2022. Pulse oximeter performance, racial inequity, and the work ahead. Respir. Care 67:252–57 [PubMed: 34772785]

66. Int. Med. Device Regulat. Forum. 2014. “Software as a medical device”: possible framework for risk categorization and corresponding considerations. https://www.imdrf.org/sites/default/files/docs/imdrf/final/technical/imdrf-tech-140918-samd-framework-risk-categorization-141013.pdf

67. Int. Med. Device Regulat. Forum. 2017. Software as a medical device (SAMD): clinical evaluation—guidance from the Food and Drug Administration Staff. US Dep. Health Hum. Serv. Food Drug Adm. Cent. Devices Radiol. Health. https://www.fda.gov/files/medical%20devices/published/Software-as-a-Medical-Device-%28SAMD%29--Clinical-Evaluation---Guidance-for-Industry-and-Food-and-Drug-Administration-Staff.pdf

68. Hardeman RR, Medina EM, Kozhimannil KB. 2016. Structural racism and supporting black lives —the role of health professionals. N. Engl. J. Med 375:2113–15 [PubMed: 27732126]

69. Hammonds EM, Reverby SM. 2019. Toward a historically informed analysis of racial health disparities since 1619. Am. J. Public Health 109:1348–49 [PubMed: 31483728]

Table 1

Mechanisms by which bias in medical technologies is introduced and perpetuated

Mechanism	Pulse oximetry	Algorithms to estimate healthcare needs	Pulmonary function testing
Physical aspects of technology design	Melanin may cause light scatter leading to performance differences based on skin pigment	Not applicable	Not applicable
Use of data that do not reflect important patient subgroups	Calibration of modulation ratio with saturation may be biased if data lack sufficient diversity	Not applicable	Not applicable
Conflation of biological and social determinants of health	Not applicable	Use of a biased outcome: healthcare costs as proxy for healthcare needs	Assumption that population-based differences in lung function represent biological differences
Evaluation and regulatory standards	Lack of robust standards to ensure equal performance in relevant subgroups	Varying standards for regulating software algorithms	Delayed appreciation of findings described in peer-reviewed medical literature