News & views

Racism in science

Forum: Medical devices

Skin colour affects oxygen-sensor accuracy

COVID-19 broadened the use of pulse oximeters for rapid blood-oxygen readings, but it also highlighted the fact that skin pigmentation alters measurements. Two groups of researchers analyse this issue, and its effects on people with dark skin.

Matthew D. Keller & Brandon Harrison-Smith
Pulse-oximetry errors affect patient outcomes

The pulse oximeter is a device that estimates a person's oxygen saturation level, a measure of the oxygen concentration in their blood, by shining light through their tissue, typically a fingertip or an earlobe (Fig. 1). As highlighted by the COVID-19 pandemic, accurate pulse-oximeter readings can be crucial for clinical decisions, especially when arterial blood-gas tests — the gold standard for determining oxygen saturation levels — are not available. But these devices give readings that are often less accurate for people who have dark skin, and this shortcoming has led to medical practices that only exacerbate the problem, making pulse oximetry emblematic of the broader issue of racial bias in medicine. The first step towards a solution must involve an orchestrated effort from those who design, use and regulate these devices.

Driven by clinical experiences early in the pandemic, Sjoding *et al.*[1] published a retrospective report showing that pulse oximeters overestimate the true oxygen saturation of Black people. This inaccuracy means that diagnoses of hypoxaemia, the condition of having low levels of oxygen in one's blood, are approximately three times more likely to be missed in Black patients than in white patients. Misdiagnosed patients are said to have occult hypoxaemia when arterial blood-gas tests indicate oxygen saturation levels of less than 88% (signalling hypoxaemia), despite pulse oximeters measuring a healthy oxygenation of more than 92%.

Since Sjoding and colleagues' report, several large retrospective studies have confirmed that darker-skinned people (those self-identifying as Black, Asian, Hispanic or a combination of these) are more likely than white people to experience occult hypoxaemia[2–5]. In one study of people with COVID-19, 35% of those self-identifying as Black had their eligibility for oxygen treatment delayed, or even missed altogether, compared with just 20% of the white people documented[2]. In another study, Black people received less therapeutic oxygen than did white people who had equivalent arterial blood-gas values[3]. A more comprehensive analysis showed that, even when baseline health conditions are taken into account, people with occult hypoxaemia are prone to organ dysfunction and in-hospital mortality, and that Black people in this group have the worst organ dysfunction[5].

Although clinical reports of skin-colour bias in pulse oximetry were not widespread until the COVID-19 pandemic, evidence for this issue has been accumulating for decades[6,7]. A comparison reported in February found that pulse-oximeter readings from nine devices were consistently less accurate for darker-skinned people than for lighter-skinned people[8]. But the study also found that testing healthy individuals under carefully controlled laboratory conditions resulted in fewer cases of occult hypoxaemia than are measured in hospitals. In fact, none of the 491 people who were tested by the authors had readings consistent with occult hypoxaemia, whereas Sjoding and colleagues tallied 187 cases out of 3,527 measurements from a cohort of 1,609 hospitalized individuals. This discrepancy highlights the need to understand how pulse-oximetry errors are exacerbated in real-world use.

All of these findings echo a long history of the health-care system using fixed racial

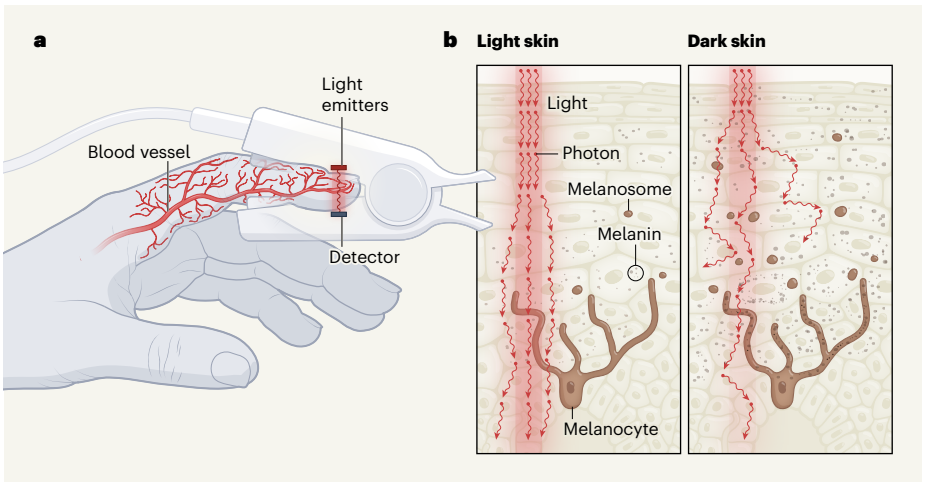


Figure 1 | Pulse-oximetry accuracy varies with skin tone. a, Devices known as pulse oximeters estimate the oxygen concentration in a person's blood by shining red and infrared light through their fingertip. Oxygenated haemoglobin absorbs infrared light more efficiently than it does red light, whereas the opposite is true for deoxygenated haemoglobin. **b**, These signals are affected by melanin, which is distributed through the skin in structures, known as melanosomes, that are produced by cells called melanocytes. Melanosomes in dark skin are both larger and more numerous than are those in light skin. Long-standing oximetry theory does not fully account for the way in which photons are scattered by the biomolecular content and structure of the tissue, and thus imprecisely corrects for the effect of pigmentation. Calibration studies compound this problem, because they typically oversample light-skinned people. This has led to overestimation of the oxygen concentration in some Black individuals' blood, and therefore to missed diagnoses of dangerously low oxygen levels.

offsets for certain instruments and risk formulas that are now recognized as potentially contributing to health inequities, rather than alleviating them[9]. For example, an algorithm that is commonly used to assess the risk of heart failure (see go.nature.com/3mw3zda) was originally designed to systematically increase the score (and thus the perceived risk) for people who are not Black. This offset came under scrutiny for raising the threshold for treating Black people, and is now an optional feature of the calculator.

In the case of pulse oximetry, the idea that race-based adjustments (rather than effective device design and calibration) could rectify the overestimation error also seems inappropriate. And although this overestimation is not solely responsible for patient-outcome disparities, such as those experienced during the COVID-19 pandemic, efforts to correct it are crucial. That's because it is increasingly clear that reports of bias in medical devices could aggravate the already-complex historical relationship between the Black community and medicine.

Sjoding and co-workers' findings prompted the US Food and Drug Administration (FDA) to issue a safety communication in February 2021 highlighting the limitations of pulse oximeters (see go.nature.com/3wkgket); it is likely that few health-care workers, and even fewer patients, had appreciated these drawbacks. And last month, the FDA announced that the Medical Devices Advisory Committee would convene in November to gather all available evidence on the issue and to determine ways of improving the accuracy of pulse oximeters. Addressing the problem appropriately will require a coordinated effort from researchers, health-care workers, device manufacturers and the FDA.

Once the mechanism of the oximetry overestimation is clearly understood, it should be possible to make this crucial piece of health-care equipment work equitably for all. This might involve altering pulse-oximeter calibration and clinical-study procedures by adopting objective metrics for skin tone, instead of using self-identified ethnicities or subjective assessments of pigmentation. An ideal solution might involve a new generation of devices that can objectively sense and account for a patient's skin tone — as well as any other factors that could affect pulse-oximetry measurements.

Matthew D. Keller is at Global Health Labs, Inc., Bellevue, Washington 98007, USA. **Brandon Harrison-Smith** is in the College of Engineering, Purdue University, West Lafayette, Indiana 47906, USA.
e-mails: matthew.keller@ghlabs.org; harr1124@purdue.edu

Chetan Patil & Mohammed Shahriar Arefin
The basis of bias in pulse oximetry

The modern finger-clip pulse oximeter was developed in the 1970s and, over the past 50 years, has revolutionized patient monitoring by enabling rapid identification of acute respiratory distress. However, both the device itself and the way it is calibrated are characterized by biases that are linked to the person's skin pigmentation. The combined consequence of these factors is an apparent racial bias in oximetry measurements that was no doubt unintended by its inventors. Overcoming these technical problems is a multifaceted challenge that requires careful analysis, and rigorous scrutiny of the way in which clinical trials are designed.

The device works by measuring the time-varying optical signal that is produced by the interaction of red and infrared light with tissue perfused with blood[10] (Fig. 1a). How light interacts with the tissue results in photons being either absorbed or scattered by molecules such as haemoglobin, melanin, lipids and water[11].

Pulse oximetry is possible because oxygenated haemoglobin absorbs infrared light more efficiently than it does red light, whereas the opposite is true for deoxygenated haemoglobin. The device shines red and infrared light through a person's skin and the detected light produces an oscillating signal, because the amount of blood in the tissue fluctuates with each heartbeat. The average value of this oscillating signal is conventionally used to indicate the total absorbance from all the biomolecules in the tissue, whereas its amplitude quantifies fluctuations in the concentration of oxygenated haemoglobin throughout the cardiac cycle.

> **"Increased pigmentation reduces the overall intensity of optical signals."**

By calculating the ratio of this amplitude to the average for red light, and normalizing it by the same ratio for infrared light, one arrives at an oximetry parameter that is linearly related to measurements of arterial blood oxygen saturation. Precise determination of this relationship for specific devices is performed through calibration studies that compare oximetry parameter values with oxygen levels in blood samples that are measured with a gas analyser.

A long-standing misconception in oximetry is that variation in the biomolecular composition of an individual — including, for example, their melanin levels — is accounted for, because the oximetry parameter is normalized by the average values of light detected by both the red and the infrared signals. This idea was supported by the results of limited theoretical analysis[12], which considered the finger to be a homogeneous absorbing material, and did not account for the fact that light scatters differently depending on its wavelength. Such scattering effects are substantial in tissues with a multilayered anatomical structure, such as those in the finger (Fig. 1b).

Computational modelling of how light interacts with tissue offers a robust theoretical framework with which to revisit the assumptions associated with the simplified conceptual frameworks used in oximetry[13]. Such studies incorporate scattering, as well as geometric factors related to tissue anatomy and sensor configuration. Simulations of pulse oximetry have shown that increased pigmentation reduces the overall intensity of optical signals, which can result in a degraded signal-to-noise ratio, and thus explain observations of increased measurement variability in dark-skinned people[14]. Other simulations have contradicted the conventional belief that the widely used calibration parameter is not affected by pigmentation, supporting empirical findings indicating that increased pigmentation decreases the normalized ratio[15].

Currently, enrolment guidance from the FDA for testing oximeter accuracy suggests that studies should involve a minimum of ten people, at least two of whom should be "darkly pigmented" (see go.nature.com/3rc1whx). However, the shade of a person's pigmentation is an inherently subjective criterion and can contribute to inconsistency in study design. Given the evidence, both from measurements and from simulations, that the parameter used for conventional oximeters is pigmentation dependent, there is reason to question the FDA guidance that only 20% of people tested in these studies must have dark skin to achieve equitable calibration. Computational studies simulating the expected outcome of calibration studies in which 20% of people are 'darkly pigmented' support the findings of retrospective clinical studies that reveal an overestimation bias in oxygen saturation measurements in Black Americans[16].

A combination of theoretical analyses and clinical findings will ultimately strengthen our understanding of challenging issues posed by pulse oximetry; these include the effects of pigmentation, as well as those of low perfusion of blood through a person's tissue, carbon monoxide poisoning and anaemia[17]. The success of pulse oximetry as a real-time low-cost tool for monitoring a person's cardiorespiratory status has led to its widespread use, and the technique's prevalence has, in turn, highlighted situations in which inaccuracies occur. Clearly, clinical findings from

the past few years provide an imperative for developing and validating oximeters without a fundamental dependence on pigmentation. These studies also highlight the importance of carefully reconsidering the enrolment criteria suggested for calibration studies, so that the skin pigmentation of test participants is evenly balanced, and determined using objective measures.

Chetan Patil and **Mohammed Shahriar Arefin** are in the College of Engineering, Temple University, Philadelphia, Pennsylvania 19122, USA.
e-mails: c.patil@temple.edu; shahriar.arefin@temple.edu

1. Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E. & Valley, T. S. *N. Engl. J. Med.* **383**, 2477–2478 (2020).
2. Fawzy, A. *et al. JAMA Intern. Med.* **182**, 730–738 (2022).
3. Gottlieb, E. R., Ziegler, J., Morley, K., Rush, B. & Celi, L. A. *JAMA Intern. Med.* **182**, 849–858 (2022).
4. Valbuena, V. S. M. *et al. BMJ* **378**, e069775 (2022).
5. Wong, A.-K. I. *et al. JAMA Netw. Open* **4**, e2131674 (2021).
6. Ries, A. L., Prewitt, L. M. & Johnson, J. J. *Chest* **96**, 287–290 (1989).
7. Feiner, J. R., Severinghaus, J. W. & Bickler, P. E. *Anesth. Analg.* **105**, S18–S23 (2007).
8. Okunlola, O. E. *et al. Respir. Care* **67**, 252–257 (2022).
9. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. *N. Engl. J. Med.* **383**, 874–882 (2020).
10. Severinghaus, J. W. *Anesth. Analg.* **105**, S1–S4 (2007).
11. Jacques, S. L. *Phys. Med. Biol.* **58**, R37 (2013).
12. Mannheimer, P. D. *Anesth. Analg.* **105**, S10–S17 (2007).
13. Wang, L., Jacques, S. L. & Zheng, L. *Comput. Methods Programs Biomed.* **47**, 131–146 (1995).
14. Chatterjee, S. & Kyriacou, P. A. *Sensors* **19**, 789 (2019).
15. Boonya-ananta, T. *et al. Sci. Rep.* **11**, 2570 (2021).
16. Arefin, M. S., Dumont, A. P. & Patil, C. A. *Proc. SPIE* **11951**, 1195103 (2022).
17. Fine, J. *et al. Biosensors* **11**, 126 (2021).

The authors declare no competing interests.

In Retrospect

Racism in science

The unseen Black faces of AI algorithms

Abeba Birhane

An audit of commercial facial-analysis tools found that dark-skinned faces are misclassified at a much higher rate than are faces from any other group. Four years on, the study is shaping research, regulation and commercial practices.

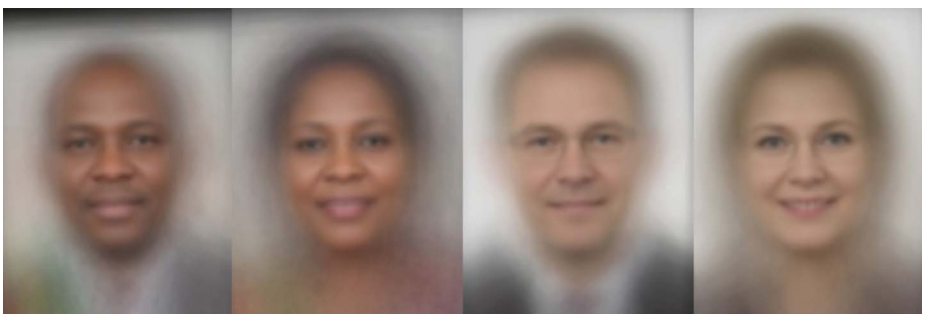
Data sets are essential for training and validating machine-learning algorithms. But these data are typically sourced from the Internet, so they encode all the stereotypes, inequalities and power asymmetries that exist in society. These biases are exacerbated by the algorithmic systems that use them, which means that the output of the systems is discriminatory by nature, and will remain problematic and potentially harmful until the data sets are audited and somehow corrected. Although this has long been the case, the first major steps towards overcoming the issue were taken only four years ago, when Joy Buolamwini and Timnit Gebru[1] published a report that kick-started sweeping changes in the ethics of artificial intelligence (AI).

As a graduate student in computer science, Buolamwini was frustrated that commercial facial-recognition systems failed to identify her face in photographs and video footage. She hypothesized that this was due, in part, to the fact that dark-skinned faces were not represented in the data sets that were used to train the computer programs she was studying. This insight led Buolamwini and her collaborator Gebru to undertake a systematic audit of commercial facial-analysis systems, and to demonstrate that such systems perform differently depending on the skin colour and gender of the person in the image. The work became known as the Gender Shades audit.

The authors began by using a skin-type classification system, approved by dermatologists, to assess the composition of two image banks, known as IJB-A and Adience, that were widely used at the time to train facial-recognition software. They found that individuals with light-coloured skin were the subject of 79.6% of the images in IJB-A and of 86.2% of those in Adience. This prompted Buolamwini and Gebru to compile their own set of images — one that offered a broader range of skin tones than did either of the existing options, as well as including similar numbers of men and women (commercial algorithms are typically not capable of dealing with non-binary classifications). To do so, they turned to photographs of politicians from countries with gender parity in their national parliaments. The resulting data set, known as the Pilot Parliaments Benchmark (Fig. 1), contains images of 1,270 individuals from Rwanda, Senegal, South Africa, Iceland, Finland and Sweden.

Buolamwini and Gebru then used their benchmark set to evaluate three commercial gender-classification systems developed by the technology companies Microsoft, Face++ and IBM. Rather than assessing the accuracy of these systems on the basis of gender or of skin type, the authors compared the performance of the classifiers on four intersectional groups that they termed darker female, darker male, lighter female and lighter male. They found that women with darker skin were the most likely to be misclassified, with a maximum classification error rate of 34.7%; by contrast, the maximum error rate for men with lighter skin was 0.8%. All three systems consistently showed poor accuracy for women with dark skin and performed substantially better on white men.

Impactful research isn't always understood and acknowledged at first glance, especially when it challenges conventional thinking. At the time of publication, Buolamwini and Gebru's paper was considered an outlier — not only in the field of computer vision (the study of how computers can be made to automate tasks



DR. JOY BUOLAMWINI (CC BY 4.0)

Figure 1 | A gender-balanced facial image bank with a range of skin tones. On realizing that dark-skinned faces were under-represented in the data sets of images that are used to train facial-recognition software, Buolamwini and Gebru[1] compiled their own data set using photographs of politicians from countries with gender parity in their national parliaments. This is a subset of 'average' faces made by blending many images from the full data set, which contains photographs of 1,270 individuals from Rwanda, Senegal, South Africa, Iceland, Finland and Sweden. Buolamwini and Gebru used their data set to show that three commercial gender-classification systems misclassified women with darker skin with an error rate that was much higher than that for men with lighter skin.