ORIGINAL CLINICAL REPORT

OPEN

Utility of Skin Tone on Pulse Oximetry in Critically Ill Patients: A Prospective Cohort Study

Sicheng Hao, MS[1]
Katelyn Dempsey, MPH[1]
João Matos, MS[1]
Christopher E. Cox, MD, MPH[1]
Veronica Rotemberg, MD, PhD[2]
Judy W. Gichoya, MD[3]
Warren Kibbe, PhD[4]
Chuan Hong, PhD[4]
An-Kwok Ian Wong, MD, PhD[1,4]

OBJECTIVE: Pulse oximetry, a ubiquitous vital sign in modern medicine, has inequitable accuracy that disproportionately affects minority Black and Hispanic patients, with associated increases in mortality, organ dysfunction, and oxygen therapy. Previous retrospective studies used self-reported race or ethnicity as a surrogate for skin tone which is believed to be the root cause of the disparity. Our objective was to determine the utility of skin tone in explaining pulse oximetry discrepancies.

DESIGN: Prospective cohort study.

SETTING: Patients were eligible if they had pulse oximetry recorded up to 5 minutes before arterial blood gas (ABG) measurements. Skin tone was measured using administered visual scales, reflectance colorimetry, and reflectance spectrophotometry.

PARTICIPANTS: Admitted hospital patients at Duke University Hospital.

INTERVENTIONS: None.

MEASUREMENTS AND MAIN RESULTS: Sao_2–Spo_2 bias, variation of bias, and accuracy root mean square, comparing pulse oximetry, and ABG measurements. Linear mixed-effects models were fitted to estimate Sao_2–Spo_2 bias while accounting for clinical confounders.

One hundred twenty-eight patients (57 Black, 56 White) with 521 ABG-pulse oximetry pairs were recruited. Skin tone data were prospectively collected using six measurement methods, generating eight measurements. The collected skin tone measurements were shown to yield differences among each other and overlap with self-reported racial groups, suggesting that skin tone could potentially provide information beyond self-reported race. Among the eight skin tone measurements in this study, and compared with self-reported race, the Monk Scale had the best relationship with differences in pulse oximetry bias (point estimate: −2.40%; 95% CI, −4.32% to −0.48%; $p = 0.01$) when comparing patients with lighter and dark skin tones.

CONCLUSIONS: We found clinical performance differences in pulse oximetry, especially in darker skin tones. Additional studies are needed to determine the relative contributions of skin tone measures and other potential factors on pulse oximetry discrepancies.

Racial and ethnic bias in pulse oximetry stands out as a quintessential health inequity, whereby the same medical devices that guide clinical decision-making may fail to function equally well for all patients (1). The reliability of pulse oximetry has been a reason for concern for decades (2–6), but it was not until the COVID-19 pandemic when a seminal paper by Sjoding et al (7) reported racial bias in pulse oximetry measurements that pulse oximetry became a health equity issue.

DOI: 10.1097/CCE.0000000000001133

KEY POINTS

Question: Can skin tone capture information beyond race to help model pulse oximetry discrepancies?

Findings: Pulse oximetry bias across races seems to persist across skin tone when measured using administered visual scales, reflectance colorimetry, or reflectance spectrophotometry. Among the eight skin tone measurements in this study, and compared with self-reported race, the Monk Scale seemed to best correlate with pulse oximetry bias when comparing patients with lighter and dark skin tones.

Meaning: Compared with self-reported race, skin tone is associated with some pulse oximetry discrepancies; we recommend using skin tone to assist the regulatory clearance of equitable pulse oximeters.

Followed by other studies (8–15), oxygen saturation measured by pulse oximetry (Spo_2) is widely reported to overestimate the "true" Sao_2, measured by arterial blood gas (ABG), disproportionately affecting Black and Hispanic patients. A seemingly small discrepancy is associated with higher rates of "hidden hypoxemia" among these patients (2, 8, 12), with associated inequities in oxygen therapies (16) and increases in mortality and organ dysfunction (8). A previous study showed the discrepancies in pulse oximeters linked to delayed oxygen and pharmacologic treatments (9).

Pulse oximeters estimate Sao_2 saturation by measuring the light absorption of oxyhemoglobin and deoxyhemoglobin in capillary blood (17, 18). Previous studies have shown that skin tone can independently affect light absorption, causing discrepant readings, especially among darker-skinned individuals (19–21). As such, previous retrospective studies share a fundamental limitation: self-reported race or ethnicity is used as a surrogate for skin tone, although the root cause of these discrepancies is believed to be skin tone (22).

In this cohort study, we prospectively collected skin tone data from critically ill patients in various body locations using different devices. We paired these data with pulse oximetry measurements, ABG, and other Electronic Health Records (EHR) data to investigate the utility of skin tone data in explaining pulse oximetry performance.

As a pilot study, our objectives were two-fold: first, to provide a framework to conduct larger clinical studies that assess the association between different skin tone measurements and pulse oximetry discrepancies; and second, to provide evidence that can support recent discussions from the Food and Drug Administration (FDA) (23–26) in pursuit of guidelines to evaluate pulse oximetry performance in a more inclusive spectrum of patients.

MATERIALS AND METHODS

This study was approved by the Duke Health institutional review board (IRB) under Pro00110842 on May 18, 2022, titled "ENCODE (mEasuring skiN Color to correct pulse Oximetry DisparitiEs)," following the American Medical Association's recommendations on health equity language and adhering to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement (27). Procedures were followed in accordance with the ethical standards of the Duke Health IRB and with the Helsinki Declaration of 1975.

Cohort Selection

Patients admitted to the emergency department, adult ICU, and surgical units at Duke University Hospital were screened. Standard-of-care ABG up to 5 minutes after a pulse oximetry measurement was required for eligibility, resulting in pulse oximetry-ABG pairs.

Exclusion criteria included unremovable fingernail polish, admission for a vascular complication (e.g., grafting or stenting), amputation, and large areas of skin discoloration where the accuracy of skin tone measurements could be affected. Pairs containing either a Sao_2 or a Spo_2 measurement out of the 70–100% range were excluded (8, 28).

Data Collection and Processing

All data and patients' consent were stored in Duke Health's Research Electronic Data Capture (REDCap; Vanderbilt University, Nashville, TN), with data processing performed in Python 3.10.

The mathematical definitions of ABG-pulse oximeter bias, variation of bias, and accuracy root mean square (A_{RMS}) are in **Supplemental Formulas 2, 3**, and **4** (http://links.lww.com/CCX/B379).

Skin Tone Measurement Methods and Locations

Three types of skin tone assessment were conducted using different devices: administered visual scales (Fitzpatrick, Monk [29], and Von Luschan scales, visible in **Supplemental Fig. 5, *A*, *B,*** and ***C***, http://links.lww.com/CCX/B379); reflectance colorimetry (Device name: "Delfin SkinColorCatch," Kuopio, Finland); and reflectance spectrophotometry (Device names: "Variable Spectro 1 Pro Bridge Set," Variable, Inc, Chattanooga, TN; and "Konica Minolta CM-700D Spectrophotometer," Tokyo, Japan).

All the skin tone data are collected within 7 days of the pulse oximetry-ABG pairs and with controlled lighting to ensure reproducibility. Using three administered scales and three color measure tools, eight different skin tone measures were collected in this study, as detailed in **Supplemental Table 2** (http://links.lww.com/CCX/B379). In increasing values, four of these eight measures (Fitzpatrick, Von Luschan, Monk, Melanin Index) progress from lightest to darkest; the other four progress from darkest to lightest. Further details are demonstrated in **Supplemental Text** and Supplemental Table 2 (http://links.lww.com/CCX/B379).

The skin tone measurements were performed across 16 body locations. We took the average of four palm locations (left and right dorsal, ventral) to better represent the pulse oximeter locations (more details in **Supplemental Methods**, http://links.lww.com/CCX/B379).

We collected all pulse oximetry-ABG pairs from a patient's hospital admission once they were considered to be in the study cohort. Pulse oximetry values and ABG panel data were merged into pulse oximetry-ABG pairs and recorded in REDCap. Demographic data were merged from the EHR system. Three race groups were defined: "Black," "Other," and "White" patients. The group "Other" captures minority patients who self-identify as Asian, American Indian/Alaskan natives, more than two races, and unknown race groups that separately represent 12% of patients. Vital signs were captured within 4 hours before the pulse oximetry-ABG pair was merged from the EHR system. Mean arterial pressure (MAP) from the arterial line was preferred when available, otherwise, cuff values were used. Laboratory test values from the previous 24 hours, relative to the pulse oximetry-ABG pair, were merged (**Table 1**).

Missingness

Missing data occurred occasionally in two skin tone measurements (Variable L* and Konica Minolta L*) due to technical issues or patient refusal.

Measurement Variability

The SD was computed across the different values of the same measurement method and location, and compared with the average SDs across all locations. Average SDs across the palm and finger locations were also computed, as depicted in **Supplemental Table 3** (http://links.lww.com/CCX/B379).

Statistical Analysis

Data analysis was performed using Python 3.10 (30), as described in the Supplemental Text (http://links.lww.com/CCX/B379), and summarized using the "tableone" package (31), in **Table 2**. For each tertile of skin tone, pulse oximetry-ABGbias, and A_{RMS} were computed, as reported in **Figure 2**. Statistical analysis was conducted in R 4.3.1 (32), using the package "nlme" for the mixed-effects analysis (33, 34).

Linear Mixed-Effects Models

Linear mixed-effects models, with patient identifiers as a random effect to account for multiple pairs, were fitted and adjusted for potential confounders (race and clinical features). Pairs with missing data were dropped for the analysis.

As a baseline, we built a model to assess the effect of self-reported race on pulse oximetry inaccuracy, adjusting for pH, Sao_2, heart rate, and MAP (**Supplemental Formula 5**, http://links.lww.com/CCX/B379). The following models, documented in **Supplemental Formula 6** (http://links.lww.com/CCX/B379), included these same covariates, as well as the skin tone variables, separate per model (eight models were built in total to assess the individual effect of each skin tone variable).

Lastly, to investigate the combined effect of all the skin tone measurements on pulse oximeter inaccuracy, we fitted two linear mixed-effects models (**Supplemental Formulas 7** and **8**, http://links.lww.com/CCX/B379). The first model included six skin tone variables, excluding Konica Minolta L* and Variable L* due to missingness.

TABLE 1.
Characteristics of Arterial Blood Gas Samples and Paired-Pulse Oximetry by Race Group

Characteristics	Sao_2–Spo_2 Pairs Grouped by Race Group				
	Missing	Black	Other	White	Overall
n		232	24	265	521
Sao_2, mean (SD)	0	95.6 (2.4)	95.9 (2.0)	95.8 (2.2)	95.7 (2.3)
Spo_2, mean (SD)	0	97.2 (3.1)	97.7 (2.2)	97.3 (2.9)	97.3 (2.9)
Sao_2–Spo_2, mean (SD)	0	−1.6 (1.9)	−1.8 (1.8)	−1.5 (2.4)	−1.6 (2.1)
First-day Sequential Organ Failure Assessment, median (Q1, Q3)	159	9.0 (7.0, 11.0)	6.0 (6.0, 8.0)	10.0 (6.0, 15.0)	10.0 (6.0, 11.0)
Sex (female), *n* (%)	0	91 (39.2)	3 (12.5)	107 (40.4)	201 (38.6)
pH, mean (SD)	39	7.4 (0.1)	7.4 (0.1)	7.4 (0.1)	7.4 (0.1)
Heart rate, mean (SD)	8	93.5 (18.9)	94.1 (15.5)	96.1 (19.6)	94.8 (19.1)
Main arterial pressure, mean (SD)	19	81.7 (17.9)	95.8 (54.1)	81.4 (33.5)	82.2 (29.3)
Sodium, mean (SD)	15	139.0 (5.3)	138.2 (2.7)	137.5 (5.7)	138.2 (5.4)
Platelet, mean (SD)	28	180.1 (102.5)	279.8 (186.4)	202.2 (125.8)	195.9 (121.3)
Potassium, mean (SD)	14	4.1 (0.6)	3.9 (0.6)	4.0 (0.5)	4.1 (0.6)
Blood urea nitrogen, mean (SD)	26	24.3 (17.4)	20.2 (11.0)	31.8 (22.7)	27.9 (20.4)
WBC, mean (SD)	44	13.9 (7.1)	17.2 (7.6)	14.5 (8.7)	14.4 (8.0)
Creatinine, mean (SD)	26	1.5 (1.0)	1.5 (1.4)	1.9 (1.4)	1.7 (1.3)
Glucose, mean (SD)	9	152.7 (54.1)	138.3 (27.6)	151.6 (58.4)	151.5 (55.4)
Bicarbonate, mean (SD)	26	24.7 (3.8)	23.7 (3.0)	23.7 (5.0)	24.1 (4.4)
Chloride, mean (SD)	26	105.3 (6.1)	105.5 (4.4)	103.8 (6.5)	104.6 (6.3)
Hemoglobin, mean (SD)	15	10.0 (1.9)	11.2 (1.3)	9.5 (1.6)	9.8 (1.8)
Norepinephrine equivalent dose (μg/kg/min), mean (SD)	0	0.1 (0.1)	0.0 (0.0)	0.1 (0.1)	0.1 (0.1)

This table depicts the pair-level characteristics of the obtained cohort, in terms of sex distribution, pulse oximetry, and arterial blood gas measurements, as well as paired laboratory test values. The latter present missingness due to the applied pairing criteria; nevertheless, covariates' missingness is always < 10% (in the 521 pairs).

The second model included all eight skin tone measurements as a sensitivity analysis. These differences in design resulted in a lower sample size for the second model. **Table 3** summarizes the built models. All the significance levels in linear mixed-effects models are calculated using likelihood ratio tests (LRTs) using chi-square statistics.

RESULTS

Cohort Characteristics

From January 1, 2023, to June 30, 2023, 1,119 inpatients with qualifying pulse oximetry-ABG pairs were screened at Duke University Hospital. Of the 302 patients who met our inclusion criteria, 134 consented to participate (**Fig. 1**). After excluding 6 patients due to withdrawal, missing location data, or incomplete data, 128 patients remained for analysis (39.8% female, 43% Black; see Table 2). From these patients, 521 pulse oximetry-ABG pairs were obtained after excluding readings outside the 70–100% range. Spo_2 values ranged from 82% to 100%, and Sao_2 values from 83.8% to 99.0%. The differences between Sao_2 and Spo_2 ranged from –9.0% to 8.8% (see Table 2 and **Supplemental Table 1** [http://links.lww.com/CCX/B379] for detailed pair-level characteristics).

Measurement Variability Across Skin Tone Locations

We assessed the consistency of skin tone measurements by comparing readings from three different examiners across three days within one week on a single volunteer. Color

TABLE 2.
Characteristics of the Cohort Obtained After Applying Inclusion and Exclusion Criteria, Grouped by Race

Characteristics	Grouped by Race				
	Missing	Black	Other	White	Overall
n		57	15	57	129
Ethnicity, *n* (%)					
Not Hispanic/Latino	0	57 (100.0)	12 (80.0)	53 (93.0)	122 (94.6)
Hispanic/Latino			1 (6.7)	3 (5.3)	4 (3.1)
Unknown			2 (13.3)	1 (1.8)	3 (2.3)
Gender, *n* (%)					
Female	0	23 (40.4)	3 (20.0)	26 (45.6)	52 (40.3)
Observed oximeter location, *n* (%)					
Finger or missing	0	54 (94.7)	14 (93.3)	57 (100.0)	125 (96.9)
Forehead		1 (1.8)			1 (0.8)
Right toe		2 (3.5)	1 (6.7)		3 (2.3)
	0	96.6 (94.8, 97.1)	97.0 (96.3, 97.3)	96.0 (94.9, 97.2)	96.6 (94.9, 97.2)
Delfin index E, mean (SD)	0	98.0 (96.0, 100.0)	98.0 (96.5, 100.0)	97.0 (95.0, 99.0)	98.0 (96.0, 100.0)
Fitzpatrick scale, mean (SD)	0	−2.1 (−2.7, −0.7)	−1.0 (−2.3, −0.5)	−1.9 (−2.6, −0.1)	−2.0 (−2.6, −0.3)
Von Luschan scale, mean (SD)	0	435.4 (15.6)	428.0 (17.2)	408.0 (18.9)	422.4 (21.6)
Monk scale, mean (SD)	0	742.2 (42.3)	650.2 (51.7)	600.3 (51.0)	668.8 (82.1)
Delfin individual typology angle, mean (SD)	0	−2.9 (13.9)	21.0 (12.5)	36.6 (12.4)	17.4 (22.8)
Delfin L*, mean (SD)	0	48.9 (4.8)	56.4 (6.6)	62.9 (4.5)	56.0 (8.2)
Variable L*, mean (SD)	0	7.4 (2.3)	6.5 (2.8)	4.8 (2.6)	6.2 (2.8)
Konica Minolta L*, mean (SD)	0	18.3 (2.5)	20.1 (3.4)	17.0 (2.8)	17.9 (2.9)
Delfin Melanin Index mean (SD)	0	4.8 (0.6)	3.6 (1.1)	2.7 (0.6)	3.7 (1.2)
Von Luschan chromatic scale, mean (SD)	0	27.3 (2.3)	22.0 (5.2)	18.9 (3.5)	22.9 (5.1)
Monk Skin Tone Scale, mean (SD)	0	6.4 (0.6)	5.2 (1.4)	4.3 (0.7)	5.3 (1.3)
First ICU, *n* (%)					
Medical ICU	7	9 (16.4)	3 (20.0)	12 (23.1)	24 (19.7)
Other ICU		16 (29.1)	3 (20.0)	10 (19.2)	29 (23.8)
Surgical ICU		30 (54.5)	9 (60.0)	30 (57.7)	69 (56.6)
ICU length of stay, median (Q1, Q3)	38	4.3 (2.0, 11.1)	2.3 (0.9, 3.1)	5.5 (2.5, 8.6)	4.1 (1.8, 9.2)
Hospital length of stay, median (Q1, Q3)	0	21.0 (9.0, 39.0)	10.0 (5.5, 13.5)	17.0 (8.0, 33.0)	17.0 (8.0, 33.0)

Demographic information for all 128 patients, along with their skin tone measurements, were grouped by race. The group "Other" contains patients who self-identify as Asian ($n = 5$), American Indian/Alaskan natives ($n = 6$), more than two races ($n = 2$), and unknown race ($n = 2$). Among the eight Skin Tone Scales, the Monk Scale, Fitzpatrick scale, Von Luschan scale, and Delfin Melanin Index are ordered numerically ascending from light to dark, the other ones are ascending. Detailed descriptions of each skin tone variable are shown in Supplemental Table 2 (http://links.lww.com/CCX/B379).

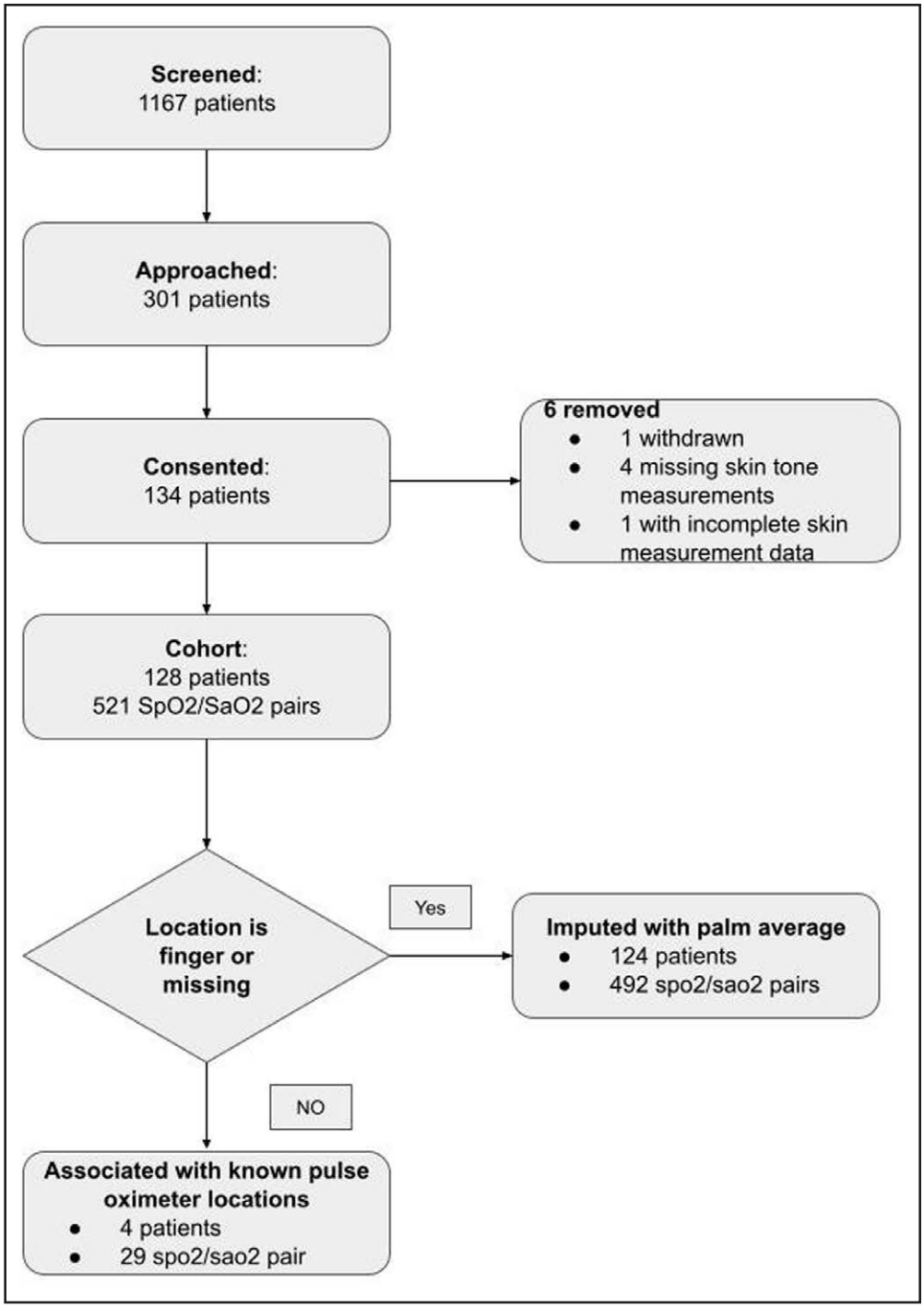

Figure 1. Flow diagram. A total of 1167 patients were screened. Exclusion criteria included unremovable fingernail polish, admission for a vascular complication (e.g., grafting or stenting), amputation, and large areas of skin discoloration where the accuracy of skin tone measurements could be affected due to arterial insufficiency or cytopenias. Pairs containing either a Sao_2 or a Spo_2 measurement of the 70–100% range were excluded. Of these, 301 patients qualified for this prospective study and were approached. Among the 134 patients who signed consent forms, one patient later withdrew, one patient did not have complete skin measurement data, and four patients did not have skin measurements. For patients who had pulse oximetry measurements done on the finger, we used the average of four palm locations (left ventral, right ventral, left dorsal, right dorsal). For patients who did not have pulse oximetry locations specified, we presumed the measurement was done on the finger and imputed it using the four palm locations as well.

measurement devices showed lower SDs than administered visual scales (Supplemental Table 3, http://links.lww.com/CCX/B379). When measuring skin tone, areas such as the palms, sternum, and underarms, were more stable (had lower variability) among different measurement sites, as opposed to fingers, toes, and earlobes. The average of left and right palms was found to be the most stable, justifying their use as the preferred skin tone measurement for analysis when the pulse oximeter was on the finger, which was 125 of 128 patients (see Supplemental Table 2, http://links.lww.com/CCX/B379, for detailed descriptions of skin tone variables).

Skin Tone Measurements by Race

Analysis showed a wide variability in skin tone measurements within racial groups, after converting all skin into the same range (0–1), with 0 being the lightest and 1 being the darkest patient's skin tone (**Supplemental Formula 1**, http://links.lww.com/CCX/B379). Black patients' skin tone ranges between 0.2 and 1, and White patients' skin tone spectrum ranges between 0 and 0.8 across eight different skin tone measurements on their pulse oximeter location. An overlap of skin tone between White and Black is observed: the middle tertiles of skin tone have approximately equal numbers

TABLE 3.
Results of the Adjusted Linear Mixed-Effects Models

Model	Variables	Coefficients	*p*	*N*	Log-Lik
Association between race Spo_2 bias (Supplemental Formula 5, http://links.lww.com/CCX/B379)	White	Baseline	0.64	463	−994.3
	Black	−0.23			
	Other	−0.31			
Association between separate Skin Tone Scale and Spo_2 bias (Supplemental Formula 6, http://links.lww.com/CCX/B379)	Fitzpatrick	−0.75	0.24	463	−993.6
	Von Luschan	−1.10	0.16	463	−993.3
	Monk	−2.40	**0.01**	463	−991.2
	Delfin individual typology angle	−0.62	0.45	463	−993.9
	Delfin L*	0.06	0.98	463	−994.3
	Konica Minolta L*	−0.77	0.38	424	−914.0
	Variable L*	−0.31	0.67	367	−784.4
	Delfin Melanin Index	0.17	0.87	463	−994.3
Association between six Skin Tone Scales and bias (Supplemental Formula 7, http://links.lww.com/CCX/B379)	Expected total effect[a]	**−1.72**	**0.02**	463	−986.8

[a]Expected total effect: the expected difference in estimated measurement bias of the darkest and lightest subject (assuming the normalized value of all skin tone measurements is 1 for the darkest subject and 0 for the lightest), computed as the sum of the separate coefficients.

Results of the four linear mixed-effects models with clinical variables (Sao_2, pH heart rate, and mean arterial pressure) adjusted (Supplemental Formulas 5–8, http://links.lww.com/CCX/B379). Likelihood ratio tests (LRTs) are performed to demonstrate whether the null hypothesis should be rejected. Variables and coefficients are derived from the linear mixed-effects model with a negative value being a larger magnitude of bias, χ^2 statistics, and *p* values are derived from LRT results. *N* is the sample size of each model. ***Red cells*** represent negative coefficient values, that is, the variable affects an overestimation of Sao_2, and vice versa for ***green cells***. ***Bold***, ***underlined p*** values denote that the significance threshold was passed at 0.05 and the null hypothesis was rejected. The self-reported race alone (Supplemental Formula 5, http://links.lww.com/CCX/B379) presents coefficients in the expected direction (−0.23%; 95% CI, −0.76 to 0.30%; $p = 0.64$ for Black patients, compared with White patients), but the *p* value is not significant. When assessing the effect of a separate Skin Tone Scale on bias (Supplemental Formula 6, http://links.lww.com/CCX/B379), only the Monk Skin Tone Scale is shown to be significant (−2.40%; 95% CI, −4.32% to −0.48%; $p = 0.01$). The effect of all combined six Skin Tone Scales on the bias (the ones without missingness, Supplemental Formula 7, http://links.lww.com/CCX/B379) was found to be significant, with an expected total effect

[a]of −1.72%, $p = 0.02$. Finally, when considering all eight Skin Tone Scale variables, this expected total effect remains in the expected direction (−3.80%), although the *p* value is not significant ($p = 0.06$).

of Black and White patients (see **Supplemental Figs. 2** and **3**, http://links.lww.com/CCX/B379).

Skin Tone Association With Pulse Oximetry-ABG Bias

Figure 2 depicts the unadjusted pulse oximetry-ABG bias, and A_{RMS} across skin tone tertiles. Bias and A_{RMS} are significant across measurement methods and skin tone tertiles. The middle tertile showed the highest degree of pulse oximetry-ABG bias, especially among the objective measurement of skin tones using devices (further details in the Supplemental Text, http://links.lww.com/CCX/B379).

A_{RMS} for all three skin tone tertiles are mostly between 2% and 3%. When compared with tertile skin tones, a decreased trend in A_{RMS} is observed when skin color moves from lighter to darker.

Linear Mixed Effect Model on Pulse Oximetry-ABG Bias

A linear mixed-effects model assessed the effects of self-reported race and skin tone on pulse oximetry discrepancies. Black and other minority patients were found to have an overestimated Spo_2 (−0.23%; 95% CI, −0.76% to 0.30% for group Black and −0.31%; 95% CI,

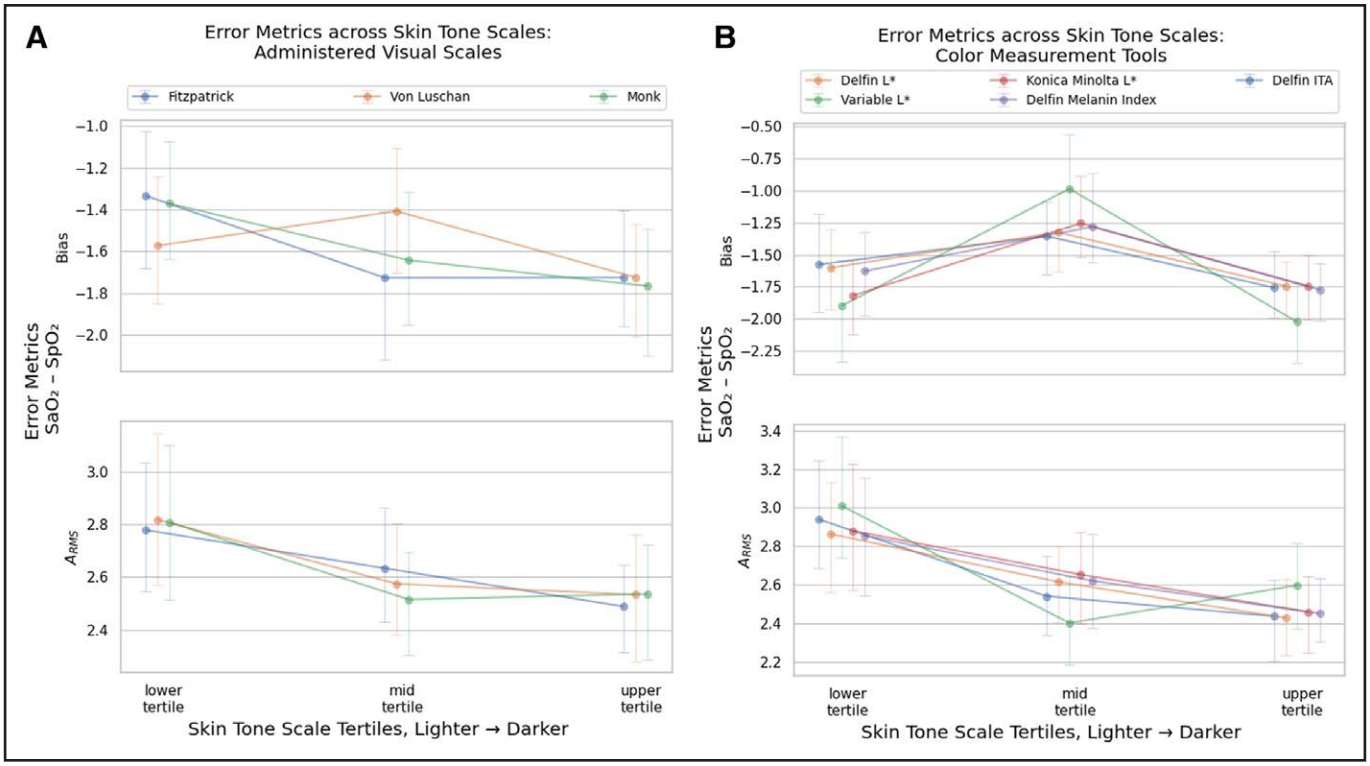

Figure 2. Bias and accuracy root mean square (A_{RMS}) across Skin Tone Scale tertiles. Unadjusted error metrics of Sao_2–Spo_2 bias and A_{RMS} across skin tone tertiles. Tertiles are ordered from lightest to darkest, from the left to the right, under each metric. For example, for the Monk Skin Tone scale, the bias is: lightest tertile: −1.371; 95% CI, −1.646 to −1.113); mid tertile: −1.643; 95% CI, −1.890 to −1.340; darkest tertile −1.767; 95% CI, −2.038 to −1.515. Other Skin Tone Scales show similar trends, whereby the lightest tertiles present a lower Sao_2–Spo_2 bias. For precision and A_{RMS}, the trends are reversed with darker tertiles presenting lower precision and A_{RMS}. Visual representations and sensitivity analysis are shown in **Supplement Figure 6** (http://links.lww.com/CCX/B379). **A**, Bias and A_{RMS} in Administered Visual Scales: Monk Skin Tone, Fitzpatrick Skin Type, Von Luschan. **B**, Bias and A_{RMS} in Color Measurement Tools. ITA = individual typology angle.

–1.31% to 0.69% for group Other) but the effect sizes did not significantly differ ($p = 0.64$). Among the eight skin tone measurements examined, only the Monk Scale was significantly associated with pulse oximetry-ABG bias (–2.40%; 95% CI, –4.32% to –0.48%; $p = 0.01$; see Table 3 for full model report).

When searching for whether a model with all six skin tone variables together could reveal a stronger association with pulse oximetry-ABG bias by examining the combined effect of six skin tone variables (excluding Konica Minolta L* and Variable L* due to missing data), the LRT rejected the null hypothesis ($p = 0.02$), indicating that at least one of these six measurements is associated with pulse oximetry-ABG bias.

DISCUSSION

The aim of this study was to explore how skin tone measurements affect pulse oximetry discrepancies. We prospectively gathered skin tone data from 128 critically ill patients, yielding 521 pairs of pulse oximetry-ABG data. We used eight different skin tone measurements across six measurement tools to quantify skin tone. Our findings show that skin tone Our findings show that skin tone measurements correlate with each other (**Supplemental Fig. 1**, http://links.lww.com/CCX/B379) and vary within racial groups, and overlaps between Black and White patients (**Supplemental Fig. 4**, http://links.lww.com/CCX/B379), indicating that skin tone data provides additional information beyond self-reported race. We also found skin tone can provide additional information beyond race when evaluating pulse oximeter discrepancies but was not enough to estimate all the variations in this study. This approach addresses a key limitation of prior studies on pulse oximetry racial discrepancies, which relied solely on racial or ethnic categories as proxies for skin tone (7–10, 12–14, 16).

Our findings show that skin tone varies within racial groups. Both Black and White racial groups contain about 80% of all Skin Tone Scales across different measurement

methods, suggesting significant overlap regardless of measurement method (Supplemental Figs. 2 and 4, http://links.lww.com/CCX/B379). Skin tone measurements also overlap between Black and White patients, the middle tertile of skin tone contains roughly equal numbers of Black and White patients. Considering skin tone measurements are either continuous or categorical with more than two categories, measuring skin tone data provides additional information beyond self-reported race.

When examining the relationship between skin tone and pulse oximetry bias, we found that racial and ethnic disparities persist (**Supplemental Table 4**, http://links.lww.com/CCX/B379), with darker-skinned patients showing a higher degree of pulse oximetry-ABG bias (Table 3 and Fig. 2). These findings are consistent with a recent report by Fawzy et al (9), who distinguished between non-Black Hispanic and Black African American patients. In contrast, models that rely solely on self-reported race without using more robust methods to ensure that Black patients have darker skin tones—a common issue where Hispanic patients are sometimes identified as Black in EHR—found the effect of race to be nonsignificant (35). Our models, which incorporate skin tone measurements, identified at least one significant skin tone measurement (Supplemental Formula 7, http://links.lww.com/CCX/B379, results in Table 3). This suggests that skin tone contributes to pulse oximetry bias independently of self-reported race.

The FDA commonly assesses pulse oximetry performance using the A_{RMS}, as detailed in Supplemental Formula 4 (http://links.lww.com/CCX/B379). A_{RMS} signifies the average deviation of Spo_2 from Sao_2. Although our linear mixed effect models find the skin tone variables have significant associations with pulse oximeter discrepancies, the remaining unexplained variations in the model measured by log-likelihood are still large. This suggests that skin tone is unlikely to be the sole contributor to performance discrepancies in pulse oximetry.

In this study, we evaluated various skin tone assessment methods and devices, ranging from low-cost options like color-printed scales to high-end devices such as Konica Minolta's spectrophotometer, which can cost thousands of dollars. Although we observed non-negligible measurement variability, it was slightly lower for device measurement skin tone, consistent with expectations. Recent FDA guidelines have proposed the use of the Monk Scale and Konica Minolta's spectrophotometer for measuring patients' skin tone (36). However, in our dataset, we did not find a skin tone measurement device or a measurement method to be significantly better performance. Therefore, we advocate for further investigation, including a broader range of skin tone measurement devices and a larger sample size, to better understand their effectiveness and potential impact. Considering the prohibitive cost of replacing existing pulse oximeters, our work stands as a fundamental milestone toward any interim solution that may address pulse oximetry inaccuracies using existing technologies (37). We propose that incorporating skin tone measurements could help mitigate residual confounding in algorithms solely reliant on self-reported race. This suggests that clinical algorithms performing a correction, rather than simple race corrections, could be more attainable.

In response to the recent discussions of the FDA on pulse oximetry performance discrepancies, we believe that this pilot study provides initial evidence to support the suggested need to thoughtfully collect and assess skin tone data in pulse oximetry clearances across multiple skin tone measurement methods (25). Besides being an important factor in pulse oximetry miscalibration, and a more objective measure than self-reported race, skin tone data seems to yield utility in pulse oximetry discrepancies. Consequently, besides requiring racial and ethnic diversity for pulse oximetry clearance, we recommend the FDA require the quantification and representation of a full spectrum of skin tones, while not disregarding the potential impact of other unmeasured confounders (26). Recognizing Beer-Lambert's law's impact on light transmission, we underline the importance of assessing other potential confounding variables such as perfusion, skin thickness, systemic vascular resistance, or local vascular resistance, for which further investigation is necessary as these are not commonly measured in medicine. Considering that bias in the direction of overestimation of Sao_2 may carry more downstream clinical harm than bias in the opposite direction, we would like to build upon previous concerns and bring to debate the question: "What is an equitable performance assessment metric for pulse oximetry for regulatory clearance?" (7).

Limitations and Future Work

Our study has several limitations and opportunities for future research. First, our skin tone data exhibited non-negligible measurement variability across sites

and examiners. To address this, we averaged the left and right side dorsal and ventral palm readings to obtain more stable skin tone measurements in 125 out of 128 patients who had pulse oximetry on their fingers. However, further refinement of skin tone measurement techniques and procedures may be necessary to improve accuracy. Second, our study was conducted at a single medical center, limiting the generalizability of our findings. Future studies with larger sample sizes and diverse patient populations across multiple institutions are needed to better characterize skin tone differences across various demographics and geographical regions. Third, although our cohort included over 40% Black patients, the representation of the darkest skin tones was limited. This may be attributed to the population's skin tone distribution in our study community. Therefore, efforts to recruit a more diverse range of skin tones should be prioritized in future studies. Additionally, we aim to enroll more patients with hypoxemia (Sao_2 < 88%) in future studies to investigate the impact of skin tone on hidden hypoxemia phenomena. Furthermore, interactions between the race and ethnicity of the rater and the patient may influence the accuracy of administered scales. Increasing the number and diversity of scale raters can help mitigate potential biases in skin tone assessment. Furthermore, we plan to examine other potential covariates that may contribute to pulse oximetry disparities, such as perfusion, skin thickness, systemic vascular resistance, or local vascular resistance. Understanding the role of these factors can provide valuable insights into pulse oximetry accuracy across diverse patient populations. Finally, although our prospective study suggests that skin tone is unlikely to be the sole contributor to pulse oximetry discrepancies, further investigation into other unmeasured confounders is warranted. By comprehensively assessing various factors influencing pulse oximetry performance, we can develop more effective strategies to improve accuracy and equity in healthcare delivery.

CONCLUSIONS

This pilot study analyzed skin tone measurements with pulse oximetry performance discrepancies among critically ill patients. We prospectively collected skin tone assessments via administered visual scales, reflectance colorimetric, and spectrophotometric devices and found using race as a proxy for skin tone measurements has limitations. Similarly to previous reports, darker-skin-toned patients yielded a greater bias, independently of clinical confounders. However, with a large variation in pulse oximetry data, skin tone is unlikely to be the sole contributor to performance discrepancies in pulse oximetry. When comparing different skin tone measurement scales, enough evidence supports which skin tone measurements is the best to capture pulse oximeter bias.

1 *Division of Pulmonary, Allergy, and Critical Care Medicine, Department of Medicine, Duke University, Durham, NC.*

2 *Dermatology Service, Memorial-Sloan Kettering Cancer Center, New York, NY.*

3 *Department of Radiology, Emory University School of Medicine, Atlanta, GA.*

4 *Division of Translational Biomedical Informatics, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC.*

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's website (http://journals.lww.com/ccejournal).

Drs. Cox and Wong participated in conceptualization. Drs. Dempsey and Hao participated in data curation. Drs. Hao, Matos, and Hong participated in the formal analysis. Dr. Wong participated in the funding acquisition. Dr. Dempsey and Wong participated in the investigation. Drs. Hong, Kibbe, and Hao participated in the methodology. Dr. Wong and Kibbe participated in supervision. Dr. Matos participated in validation. Drs. Matos and Hao participated in visualization. Drs. Dempsey, Hao, and Matos participated in writing the original draft. Drs. Cox, Rotemberg, Gichoya, Hong, and Wong participated in reviewing and editing the writing.

Dr. Wong holds equity and management roles in Ataia Medical. Dr. Wong is supported by the Duke CTSI by the National Center for Advancing Translational Sciences of the National Institutes of Health under UL1TR002553 and REACH Equity under the National Institute on Minority Health and Health Disparities of the National Institutes of Health under U54MD012530. Dr. Gichoya is a 2022 Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program and declares support from the RSNA Health Disparities grant (Number EIHD2204), Lacuna Fund (Number 67), Gordon and Betty Moore Foundation, and National Institutes of Health (National Institute of Biomedical Imaging and Bioengineering) Medical Imaging and Data Resource Center grant under contracts 75N92020C00008 and 75N92020C00021.

Drs. Hao and Dempsey have shared co-first authors.

For information regarding this article, E-mail: med@aiwong.com

REFERENCES

1. Charpignon M-L, Byers J, Cabral S, et al: Critical bias in critical care devices. *Crit Care Clin* 2023; 39:795–813

2. Jubran A, Tobin MJ: Reliability of pulse oximetry in titrating supplemental oxygen therapy in ventilator-dependent patients. *Chest* 1990; 97:1420–1425
3. Nickerson BG, Sarkisian C, Tremper K: Bias and precision of pulse oximeters and arterial oximeters. *Chest* 1988; 93:515–517
4. Perkins GD, McAuley DF, Giles S, et al: Do changes in pulse oximeter oxygen saturation predict equivalent changes in arterial oxygen saturation? *Crit Care* 2003; 7:R67
5. Singh AK, Sahi MS, Mahawar B, et al: Comparative evaluation of accuracy of pulse oximeters and factors affecting their performance in a tertiary intensive care unit. *J Clin Diagn Res* 2017; 11:OC05–OC08
6. Ross PA, Newth CJL, Khemani RG: Accuracy of pulse oximetry in children. *Pediatrics* 2014; 133:22–29
7. Sjoding MW, Dickson RP, Iwashyna TJ, et al: Racial bias in pulse oximetry measurement. *N Engl J Med* 2020; 383:2477–2478
8. Wong A, Charpignon M, Kim H, et al: Analysis of discrepancies between pulse oximetry and arterial oxygen saturation measurements by race and ethnicity and association with organ dysfunction and mortality. *JAMA Netw Open* 2021; 4:e2131674
9. Fawzy A, Wu TD, Wang K, et al: Racial and ethnic discrepancy in pulse oximetry and delayed identification of treatment eligibility among patients with COVID-19. *JAMA Intern Med* 2022; 182:730–738
10. Valbuena VSM, Seelye S, Sjoding MW, et al: Racial bias and reproducibility in pulse oximetry among medical and surgical inpatients in general care in the Veterans Health Administration 2013-19: Multicenter, retrospective cohort study. *BMJ* 2022; 378:e069775
11. Valbuena VSM, Barbaro RP, Claar D, et al: Racial bias in pulse oximetry measurement among patients about to undergo extracorporeal membrane oxygenation in 2019-2020: A retrospective cohort study. *Chest* 2022; 161:971–978
12. Henry NR, Hanson AC, Schulte PJ, et al: Disparities in hypoxemia detection by pulse oximetry across self-identified racial groups and associations with clinical outcomes. *Crit Care Med* 2022; 50:204–211
13. Jamali H, Castillo LT, Morgan CC, et al: Racial disparity in oxygen saturation measurements by pulse oximetry: Evidence and implications. *Ann Am Thorac Soc* 2022; 19:1951–1964
14. Chesley CF, Lane-Fall MB, Panchanadam V, et al: Racial disparities in occult hypoxemia and clinically based mitigation strategies to apply in advance of technological advancements. *Respir Care* 2022; 67:1499–1507
15. Ward E, Katz MH: Confronting the clinical implications of racial and ethnic discrepancy in pulse oximetry. *JAMA Intern Med* 2022; 182:858
16. Gottlieb ER, Ziegler J, Morley K, et al: Assessment of racial and ethnic differences in oxygen supplementation among patients in the intensive care unit. *JAMA Intern Med* 2022; 182:849–858
17. Chan ED, Chan MM, Chan MM: Pulse oximetry: Understanding its basic principles facilitates appreciation of its limitations. *Respir Med* 2013; 107:789–799
18. Kirson LE, Koltes-Edwards R: Pulse oximetry. *Anesth Secrets E-Book* 2010; 168. Available at: https://books.google.com/books?hl=en&lr=&id=D6j3oydmS3AC&oi=fnd&pg=PA168&dq=hidden+hypoxemia+pulse+oximetry&ots=1MPTxADsds&sig=RWGzhrcidEHX-6DWd93PcrYqWJs. Accessed August 3, 2024
19. Bickler PE, Feiner JR, Severinghaus JW: Effects of skin pigmentation on pulse oximeter accuracy at low saturation. *Anesthesiology* 2005; 102:715–719
20. Feiner JR, Severinghaus JW, Bickler PE: Dark skin decreases the accuracy of pulse oximeters at low oxygen saturation: The effects of oximeter probe type and gender. *Anesth Analg* 2007; 105:S18–S23
21. Shi C, Goodall M, Dumville J, et al: The accuracy of pulse oximetry in measuring oxygen saturation by levels of skin pigmentation: A systematic review and meta-analysis. *BMC Med* 2022; 20:267
22. Holder AL, Wong A-KI: The big consequences of small discrepancies: Why racial differences in pulse oximetry errors matter. *Crit Care Med* 2022; 50:335–337
23. U.S. Food and Drug Administration: Center for Devices and Radiological Health: Pulse oximeter accuracy and limitations: FDA safety communication. 2023. Available at: https://public4.pagefreezer.com/content/FDA/20-02-2024T15:13/https://www.fda.gov/medical-devices/safety-communications/pulse-oximeter-accuracy-and-limitations-fda-safety-communication. Accessed August 3, 2024
24. U.S. Food and Drug Administration: Center for Devices and Radiological Health: CDRH takes steps to advance further discussions on pulse oximeters. 2023. Available at: https://www.fda.gov/medical-devices/medical-devices-news-and-events/cdrh-takes-steps-advance-further-discussions-pulse-oximeters. Accessed February 7, 2024
25. U.S. Food and Drug Administration: February 2, 2024: Anesthesiology and respiratory therapy devices panel. 2024. Available at: https://www.fda.gov/advisory-committees/advisory-committee-calendar/february-2-2024-anesthesiology-and-respiratory-therapy-devices-panel-medical-devices-advisory. Accessed February 7, 2024
26. Howard J: FDA panel recommends more diversity in pulse oximeter trials. CNN. 2024. Available at: https://www.cnn.com/2024/02/02/health/pulse-oximeters-skin-color-fda/index.html. Accessed February 7, 2024
27. Flanagin A, Frey T, Christiansen SL; AMA Manual of Style Committee: Updated guidance on the reporting of race and ethnicity in medical and science journals. *JAMA* 2021; 326:621–627
28. U.S. Food and Drug Administration: Center for Devices and Radiological Health: Pulse Oximeters—Premarket Notification Submissions [510(k)s]: Guidance for Industry and Food and Drug Administration Staff. 2020. Available at: https://www.fda.gov/regulatory-information/search-fda-guidance-documents/pulse-oximeters-premarket-notification-submissions-510ks-guidance-industry-and-food-and-drug. Accessed February 7, 2024
29. Monk E: The Monk Skin Tone Scale. 2023. Available at: https://skintone.google. Accessed August 3, 2024
30. Van Rossum G, Drake FL: Python 3 reference manual: (Python Documentation Manual Part 2). CreateSpace Independent Publishing Platform. 2009. Available at: https://play.google.com/store/books/details?id=KIybQQAACAAJ. Accessed March 10, 2024

31. Pollard TJ, Johnson AEW, Raffa JD, et al: tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open* 2018; 1:26–31
32. R Core Team: R: A Language and Environment for Statistical Computing. Vienna, R Foundation for Statistical Computing, 2020
33. Pinheiro J, Bates D: R Core Team: nlme: Linear and nonlinear mixed effects models. 2023. Available at: https://CRAN.R-project.org/package=nlme. Accessed March 10, 2024
34. Pinheiro JC, Bates DM (Eds). Linear mixed-effects models: Basic concepts and examples. In: *Mixed-Effects Models in S and S-PLUS*. New York, NY, Springer New York, 2000, pp 3–56
35. Wiles MD, El-Nayal A, Elton G, et al: The effect of patient ethnicity on the accuracy of peripheral pulse oximetry in patients with COVID-19 pneumonitis: A single-centre, retrospective analysis. *Anaesthesia* 2022; 77:143–152
36. U.S. Food and Drug Administration: Center for Devices and Radiological Health: Pulse oximeters. 2024. Available at: https://www.fda.gov/medical-devices/products-and-medical-procedures/pulse-oximeters. Accessed May 31, 2024
37. Dempsey K, Lindsay M, Tcheng JE, et al: The high price of equity in pulse oximetry: A cost evaluation and need for interim solution. *medRxiv* 2023:2023.09.21.23295939v1