medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Utility of skin tone on pulse oximetry in critically ill patients: a prospective cohort study

1. Sicheng Hao*, MS [1]
2. Katelyn Dempsey*, MPH [1]
3. João Matos, MS [1]
4. Christopher E. Cox, MD, MPH [1]
5. Veronica Rotemberg, MD, PhD [3]
6. Judy W. Gichoya, MD [4]
7. Warren Kibbe, PhD [2]
8. Chuan Hong, PhD [2]
9. Ian Wong, MD, PhD [1,2]

* co-first authors

Affiliations

1. Duke University, Department of Medicine, Division of Pulmonary, Allergy, and Critical Care Medicine, Durham, NC, USA
2. Duke University, Department of Biostatistics and Bioinformatics, Division of Translational Biomedical Informatics, Durham, NC, USA
3. Memorial-Sloan Kettering, Dermatology Service, New York, NY, USA
4. Emory University School of Medicine, Department of Radiology, Atlanta, USA

Word count: 2863

Corresponding author:
A. Ian Wong, MD, PhD
med@aiwong.com
ORCID: 0000-0001-5668-4251
Assistant Professor
Department of Medicine, Division of Pulmonary, Allergy, and Critical Care Medicine
Department of Biostatistics and Bioinformatics, Division of Translational Biomedical Informatics
Duke University
2 Genome Court, Box 103000, Durham, NC 27710
+1 (919).660.5252

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Conflicts of interest

AIW holds equity and management roles in Ataia Medical.

Funding

AIW is supported by the Duke CTSI by the National Center for Advancing Translational Sciences (NCATS) of the National Institutes of Health under UL1TR002553 and REACH Equity under the National Institute on Minority Health and Health Disparities (NIMHD) of the National Institutes of Health under U54MD012530.

JWG is a 2022 Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program and declares support from RSNA Health Disparities grant (#EIHD2204), Lacuna Fund (#67), Gordon and Betty Moore Foundation, and NIH (NIBIB) MIDRC grant under contracts 75N92020C00008 and 75N92020C00021

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Key Points

Question:
Can skin tone capture information beyond race to help explain 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 to 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 to 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.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Importance

Pulse oximetry, a ubiquitous vital sign in modern medicine, has inequitable accuracy that disproportionately affects Black and Hispanic patients, with associated increases in mortality, organ dysfunction, and oxygen therapy. Although the root cause of these clinical performance discrepancies is believed to be skin tone, previous retrospective studies used self-reported race or ethnicity as a surrogate for skin tone.

Objective

To determine the utility of objectively measured skin tone in explaining pulse oximetry discrepancies.

Design, Setting, and Participants

Admitted hospital patients at Duke University Hospital were eligible for this prospective cohort study if they had pulse oximetry recorded up to 5 minutes prior to arterial blood gas (ABG) measurements. Skin tone was measured across sixteen body locations using administered visual scales (Fitzpatrick Skin Type, Monk Skin Tone, and Von Luschan), reflectance colorimetry (Delfin SkinColorCatch [L*, individual typology angle {ITA}, Melanin Index {MI}]), and reflectance spectrophotometry (Konica Minolta CM-700D [L*], Variable Spectro 1 [L*]).

Main Outcomes and Measures

Mean directional bias, variability of bias, and accuracy root mean square (A_{RMS}), comparing pulse oximetry and ABG measurements. Linear mixed-effects models were fitted to estimate mean directional bias while accounting for clinical confounders.

Results

128 patients (57 Black, 56 White) with 521 ABG–pulse oximetry pairs were recruited, none with hidden hypoxemia. Skin tone data was prospectively collected using 6 measurement methods, generating 8 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 to self-reported race, the Monk Scale had the best relationship with differences in pulse oximetry bias (point estimate: -2.40%; 95% CI: -4.32%, -0.48%; *p*=0.01) when comparing patients with lighter and dark skin tones.

Conclusions and relevance

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.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Background

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 Sjoding and colleagues' seminal paper reported racial bias in pulse oximetry measurements that pulse oximetry became a health equity issue.[7]

Followed by other studies,[8–15] oxygen saturation measured by pulse oximetry (SpO_2) is widely reported to overestimate the "true" arterial oxygen (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 [9,16] and increases in mortality and organ dysfunction.[8]

Pulse oximeters estimate arterial oxygen saturation by measuring 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 this 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 2-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.

Pulse oximetry performance is commonly assessed by the FDA using the accuracy root mean square (A_{RMS}), with a threshold of $A_{RMS} \leq 3.0\%$ for transmittance pulse oximeters between SaO_2 and SpO_2 measurements ranging from 70–100%.[27] A_{RMS} can be decomposed into the average difference in magnitude between SaO_2 and SpO_2, often referred to as "bias", and the variability of those differences, which can be measured as the standard deviation of the difference between SaO_2 and SpO_2.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Methods

This study was approved by the Duke Health IRB under Pro00110842, following the American Medical Association's recommendations on health equity language and adhering to the STROBE statement.[28,29]

Cohort Selection

Patients admitted to an adult intensive care unit (ICU) 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 SaO_2–SpO_2 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 27]

Data Collection and Processing

All data and patients' consent were stored in Duke Health's REDCap, with data processing performed in Python 3.10.

The mathematical definitions of mean directional bias, variability of bias, and A_{RMS} are in Supplemental Formulas 2, 3, and 4.

Skin Tone

Three types of skin tone assessment were conducted using different devices: administered visual scales (Fitzpatrick, Monk[30], and Von Luschan scales, visible in Supplemental Figures 6a, 6b and 6c); reflectance colorimetry (Delfin SkinColorCatch); and reflectance spectrophotometry (Variable Spectro 1 Pro Bridge Set and Konica Minolta CM-700D Spectrophotometer).

All the skin tone data are collected within 7 days of the SaO_2–SpO_2 pairs and with controlled lighting to ensure reproducibility. Using the L*, individual typology angle (ITA) and Melanin Index (Melanin Index) color spaces, eight different skin tone measures were collected in this study, as detailed in Supplemental Table 2. Further details are demonstrated in Supplemental Text.

Data merging

Pulse oximetry values and ABG panel data were merged into SaO_2-SpO_2 pairs and recorded in REDCap. Demographic data was merged from the EHR system. Three race groups were defined, including "Black", "Other" and "White" patients. The group "Other" captures minority

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

patients who self-identify as Asian, American Indian / Alaskan natives, more than two races, and unknown race–groups that separately represent a small proportion of patients. Vital signs captured within 4 hours prior to the SaO_2–SpO_2 pair were 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 SaO_2–SpO_2 pair, were merged (listed in Supplemental 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. Missingness rates for skin tone measurements can be found in Table 1. In the merged EHR clinical data, missingness occurred in vital signs and laboratory test values when no value was found within the set windows. Missingness rates for these covariates are in Supplemental Table 1. Patients with missing data that would be necessary for modeling were dropped from the study, and sensitivity analyses were run to assess the robustness of this design choice.

Measurement variability

The standard deviation was computed across the different values of the same measure and location, and compared with the average standard deviations across all locations. Average standard deviations across palm and finger locations were also computed, as depicted in Supplemental Table 3.

Statistical Analysis

Exploratory data analysis was performed using Python 3.10[31], as described in the Supplemental Text, and summarized using the *tableone* package,[32] in Table 1. For each tertile of skin tone, mean directional bias, variability of bias, and A_{RMS} were computed, as reported in Figure 2. Statistical analysis was conducted in R 4.3.1,[33] using the packages *nlme* for the mixed-effects analysis.[34,35]

Linear mixed-effects models

Linear mixed-effects models with patient identifiers as a random effect 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 in pulse oximetry mean directional bias, adjusting for pH, SaO_2, heart rate (HR), and MAP (Supplemental Formula 5). The following models, documented in Supplemental Formula 6, 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 mean directional bias, we fitted two linear mixed-effects models (Supplemental Formula 7 and Supplemental Formula 8). The first model included six skin tone variables, excluding Konica

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Minolta L* and Variable L* due to missingness. 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 2 summarizes the built models. All the significance levels in linear mixed-effects models are calculated using likelihood ratio tests (LRT) using chi-squared statistics.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Results

Cohort Characteristics

From January 1, 2023 to June 30, 2023, a total of 1,167 admitted inpatients with qualifying SaO_2-SpO_2 pairs were screened at Duke University Hospital. Out of 301 patients who met our inclusion criteria and were approached, 134 patients consented to this study (see the flow diagram in Figure 1). After 6 exclusions due to withdrawal, missing location data, or incomplete data, 128 patients were considered for analysis (39.8% female, 43% Black, see Table 1).

After excluding readings that did not fall into the 70–100% range, a total of 521 SaO_2–SpO_2 pairs were obtained from this cohort. SpO_2 values ranged from 82% to 100%, and SaO_2 values ranged from 83.8% to 99.0%. The difference between SaO_2 and SpO_2 ranged from -9.0% to 8.8% – see Supplemental Table 1 for further pair-level characteristics.

Measurement variability on skin tone scales

Supplemental Table 3 shows that objective scales resulted in lower standard deviations when compared to subjective scales. Palm averages are found to be more stable, when compared to other locations, supporting the design choice of taking palm averages as the preferred measurement for subsequent analyses.

Skin tone and race data

Exploratory data analysis showed that the skin tone measurements yielded differences among each other and overlap with self-reported racial groups. Figure 2 depicts the unadjusted mean directional bias, variability of bias, and A_{RMS}, per skin tone tertile. Further detailed text can be found in the Supplemental Text.

Linear mixed effect model on mean directional bias

To understand the effect of self-reported race and each of the skin tones on pulse oximetry discrepancy, the fitted linear mixed-effects model did not find race to have a significant effect ($\chi^2 = 0.90$, $p = 0.64$), Nevertheless, the obtained coefficient is in the expected direction (-0.23%, 95% CI: -0.76%, 0.30% for Black patients). Among eight skin tone measurements, only the Monk scale yielded a significant effect on the mean directional bias (point estimate: -2.40%; 95% CI: -4.32%, -0.48%; p-value: 0.01). – see Table 2 for a full report of the models.

To examine the combined effect of the six skin tone variables, excluding Konica Minolta L* and Variable L* due to missing data (the goal being to maximize sample size), the LRT (Supplemental Formula 7) rejected the null hypothesis ($\chi^2 = 14.98$, $p = 0.02$). This suggests that at least one of these six skin tone measurements affects the mean directional bias.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

However, in a sensitivity analysis that included all eight skin tone variables, the LRT (Supplemental Formula 8) showed borderline insignificance ($\chi^2 = 14.86$, $p = 0.06$)

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Discussion

The objective of this study was to investigate the relationship between skin tone measurements and pulse oximetry discrepancies. We prospectively collected skin tone data from 128 critically-ill patients, comprising a total of 521 pairs of pulse oximetry-ABG data, and leveraging six different tools across sixteen body sites. We addressed the fundamental limitation of previous studies on pulse oximetry racial discrepancies,[7–10,12–14,16] which solely relied on racial or ethnic groups as a proxy for skin tone. As a prospective cohort study where we enrolled patients in a comprehensive screening process, we minimized assumptions associated with secondary data analysis[36] and obtained a cohort with equal representation of Black and White patients.

The collected skin tone measurements were shown to yield differences among scales and when compared to self-reported race, suggesting that skin tone data carries information beyond self-reported race. When assessing the relation of skin tone with pulse oximetry bias, racial and ethnic disparities[7] do seem to persist, whereby darker patients show a higher degree of mean directional bias (see Figure 2 and Table 2). These findings are aligned with a recent similar report.[37] As opposed to a model that solely relies on self-reported race (and clinical confounders) to account for pulse oximetry bias – where the effect of self-reported race is found to be nonsignificant – models that accounted for skin tone found at least one of the measures to be significant (Supplemental Formula 7, results in Table 2). This finding suggests that skin tone is related, beyond self-reported race, to pulse oximetry bias.

In this study, we tested different devices for skin tone assessment that ranged from a negligible cost (color-printed scales) to thousands of dollars (Konica Minolta's spectrophotometer). The measurement variability was non-negligible, but lower for objective scales, as expected. However, in exploring pulse oximetry bias, the Monk scale – designed to be easy-to-use for evaluation of technology, while representing a broad range of skin tones [30,38] – was found to yield the strongest association with pulse oximetry bias (Table 2). As this study does not show clear evidence that more sophisticated and expensive devices (colorimetry or spectrophotometry) add value in this application of skin tone characterization, we believe that further investigation is necessary, including a broader range of skin tone measurement devices and a larger sample size.

In the exploratory analysis, both across self-reported race and skin tone measurements, darker-pigmented patients observed a lower SaO_2–SpO_2 variability (see Supplemental Table 4 and Figure 2). To assess precision while adjusting for confounding, we identified the within-*stratum* variation by modeling heterogeneous variance in the linear mixed-effects models (Supplemental Formula 9) using the tertiles defined in Figure 2. Supplemental Table 6 lists the built models, where the within-*stratum* variance is shown to be consistently lower in the darkest tertile of all eight skin tone measurements, and higher in the lightest tertile, except for Delfin Melanin Index. Consequently, the darker-pigmented patients of our cohort seem to have more consistently wrong pulse oximetry readings, despite having a lower A_{RMS}. Although this finding is not aligned with a recent report,[37] concerns about the reliability of A_{RMS} among different racial groups in pulse oximetry performance evaluation and potential inequitable treatment delay had already

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

been raised before.[39] Considering the far-reaching implications of these findings in pulse oximetry regulation, further investigation is necessary.

Considering the prohibitive cost of replacing existing pulse oximeters,[40] our work stands as a fundamental milestone to any interim solution that may tackle pulse oximetry inaccuracies leveraging existing technologies. If the finding that pulse oximeters are "more consistently wrong" among darker-skinned patients stands in follow-up studies, one could argue that clinical algorithms that perform a holistic correction – and not simple race corrections[41] – could be more attainable, due to the observed lower variance among darker-skinned patients. Given this finding, which suggests that skin tone may provide additional information beyond race, we propose that incorporating skin tone measurements could help mitigate residual confounding in algorithms solely reliant on self-reported race.

Implications in regulation and pulse oximetry clearance

In response to FDA's recent discussions on pulse oximetry performance discrepancies,[23–25] we believe that this pilot study provides initial evidence to support the suggested need of thoughtfully collecting and assessing skin tone data in pulse oximetry clearances. 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,[26] we recommend the FDA to require the quantification and representation of a full spectrum of skin tones, while not disregarding the potential impact of other unmeasured confounders. 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.

Moreover, our findings on increased variability of bias and better A_{RMS} among darker-skinned patients, despite worsened mean directional bias, raise questions about FDA's conventional reliance on A_{RMS} for pulse oximetry clearance. These considerations require larger follow-up studies and are not necessarily aligned with other reports.[37] 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[39] and bring to debate the question: "What is an equitable performance assessment metric for pulse oximetry clearance?".

Limitations and Future Work

As our skin tone data presented non-negligible measurement variability across sites and examiners, we considered the average of the left and right, dorsal and ventral palm readings in 125 patients, out of 128, to obtain more stable skin tone measurements. Due to our limited sample size, most of the findings are not significant despite a reasonable effect size. In the future, we would like to run a larger study with more patients. Although this study's cohort had over 40% Black patients, the darkest skin tones are rarely observed, which might be due to the population skin tone distribution of the community where our study was based. Moreover, we would like to enroll more patients with hypoxemia (i.e., $SaO_2 < 88\%$), to potentially investigate

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

the impact of skin tone in hidden hypoxemia phenomena. Additionally, we would like to examine other potential covariates that may contribute to pulse oximetry disparities. Finally, as our prospective study suggests that, despite its effect, skin tone is unlikely to be the sole contributor to pulse oximetry discrepancies, we advocate for the need of further investigation on other unmeasured confounders.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Conclusion

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. Pulse oximetry varied across skin pigmentation and, 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. While these findings necessitate a larger sample size to be further validated and select the best method(s) for skin tone measurement, we hope this paper provides a framework for future similar studies, as well as initial evidence to support FDA's discussions on regulation changes towards more equitable pulse oximeters.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

References

1. Charpignon ML, Byers J, Cabral S, et al. Critical Bias in Critical Care Devices. *Crit Care Clin*. 2023;39(4):795-813. doi:10.1016/j.ccc.2023.02.005

2. Jubran A, Tobin MJ. Reliability of pulse oximetry in titrating supplemental oxygen therapy in ventilator-dependent patients. *Chest*. 1990;97(6):1420-1425. doi:10.1378/chest.97.6.1420

3. Nickerson BG, Sarkisian C, Tremper K. Bias and precision of pulse oximeters and arterial oximeters. *Chest*. 1988;93(3):515-517. doi:10.1378/chest.93.3.515

4. Perkins GD, McAuley DF, Giles S, Routledge H, Gao F. Do changes in pulse oximeter oxygen saturation predict equivalent changes in arterial oxygen saturation? *Crit Care*. 2003;7(4):R67. doi:10.1186/cc2339

5. Singh AK, Sahi MS, Mahawar B, Rajpurohit S. Comparative Evaluation of Accuracy of Pulse Oximeters and Factors Affecting Their Performance in a Tertiary Intensive Care Unit. *J Clin Diagn Res*. 2017;11(6):OC05-OC08. doi:10.7860/JCDR/2017/24640.9961

6. Ross PA, Newth CJL, Khemani RG. Accuracy of pulse oximetry in children. *Pediatrics*. 2014;133(1):22-29. doi:10.1542/peds.2013-1760

7. Sjoding MW, Dickson RP, Iwashyna TJ, Gay SE, Valley TS. Racial Bias in Pulse Oximetry Measurement. *N Engl J Med*. 2020;383(25):2477-2478. doi:10.1056/NEJMc2029240

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. doi:10.1001/jamanetworkopen.2021.31674

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(7):730-738. doi:10.1001/jamainternmed.2022.1906

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*. Published online 2022:e069775. doi:10.1136/bmj-2021-069775

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(4):971-978. doi:10.1016/j.chest.2021.09.025

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(2):204-211. doi:10.1097/CCM.0000000000005394

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(12):1951-1964. doi:10.1513/AnnalsATS.202203-270CME

14. Chesley CF, Lane-Fall MB, Panchanadam V, et al. Racial Disparities in Occult Hypoxemia

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

and Clinically Based Mitigation Strategies to Apply in Advance of Technological Advancements. *Respir Care*. 2022;67(12):1499-1507. doi:10.4187/respcare.09769

15. Ward E, Katz MH. Confronting the Clinical Implications of Racial and Ethnic Discrepancy in Pulse Oximetry. *JAMA Intern Med*. 2022;182(8):858. doi:10.1001/jamainternmed.2022.2581

16. Gottlieb ER, Ziegler J, Morley K, Rush B, Celi LA. Assessment of Racial and Ethnic Differences in Oxygen Supplementation Among Patients in the Intensive Care Unit. *JAMA Intern Med*. 2022;182(8):849-858. doi:10.1001/jamainternmed.2022.2587

17. Chan ED, Chan MM, Chan MM. Pulse oximetry: understanding its basic principles facilitates appreciation of its limitations. *Respir Med*. 2013;107(6):789-799. doi:10.1016/j.rmed.2013.02.004

18. Kirson LE, Koltes-Edwards R. Pulse oximetry. *Anesthesia Secrets E-Book*. Published online 2010:168. https://books.google.com/books?hl=en&lr=&id=D6j3oydmS3AC&oi=fnd&pg=PA168&dq=hidden+hypoxemia+pulse+oximetry&ots=1MPTxADsds&sig=RWGzhrcidEHX-6DWd93PcrYqWJs

19. Bickler PE, Feiner JR, Severinghaus JW. Effects of skin pigmentation on pulse oximeter accuracy at low saturation. *Anesthesiology*. 2005;102(4):715-719. doi:10.1097/00000542-200504000-00004

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(6 Suppl):S18-S23, tables of contents. doi:10.1213/01.ane.0000285988.35174.d9

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(1):267. doi:10.1186/s12916-022-02452-8

22. Holder AL, Wong AKI. The Big Consequences of Small Discrepancies: Why Racial Differences in Pulse Oximetry Errors Matter. *Crit Care Med*. 2022;50(2):335-337. doi:10.1097/CCM.0000000000005447

23. Center for Devices, Radiological Health. Pulse Oximeter Accuracy and Limitations: FDA Safety Communication. U.S. Food and Drug Administration. Published November 17, 2023. Accessed February 7, 2024. https://www.fda.gov/medical-devices/safety-communications/pulse-oximeter-accuracy-and-limitations-fda-safety-communication

24. Center for Devices, Radiological Health. CDRH Takes Steps to Advance Further Discussions on Pulse Oximeters. U.S. Food and Drug Administration. Published November 17, 2023. Accessed February 7, 2024. https://www.fda.gov/medical-devices/medical-devices-news-and-events/cdrh-takes-steps-advance-further-discussions-pulse-oximeters

25. February 2, 2024: Anesthesiology and Respiratory Therapy Devices Panel. U.S. Food and Drug Administration. Published January 31, 2024. Accessed February 7, 2024. https://www.fda.gov/advisory-committees/advisory-committee-calendar/february-2-2024-anesthesiology-and-respiratory-therapy-devices-panel-medical-devices-advisory

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

26. FDA panel recommends more diversity in pulse oximeter trials. *CNN*. Published online February 2, 2024. Accessed February 7, 2024. https://www.cnn.com/2024/02/02/health/pulse-oximeters-skin-color-fda/index.html

27. Center for Devices, Radiological Health. Pulse Oximeters - Premarket Notification Submissions [510(k)s]: Guidance for Industry and Food and Drug Administration Staff. U.S. Food and Drug Administration. Published February 28, 2020. Accessed February 7, 2024. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/pulse-oximeters-premarket-notification-submissions-510ks-guidance-industry-and-food-and-drug

28. 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(7):621-627. doi:10.1001/jama.2021.13304

29. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *Int J Surg*. 2014;12(12):1495-1499. doi:10.1016/j.ijsu.2014.07.013

30. Monk E. The Monk Skin Tone Scale. Published online May 2023. doi:10.31235/osf.io/pdf4c

31. Van Rossum G, Drake FL. *Python 3 Reference Manual: (Python Documentation Manual Part 2)*. CreateSpace Independent Publishing Platform; 2009. https://play.google.com/store/books/details?id=KIybQQAACAAJ

32. Pollard TJ, Johnson AEW, Raffa JD, Mark RG. tableone: An open source Python package for producing summary statistics for research papers. *JAMIA Open*. 2018;1(1):26-31. doi:10.1093/jamiaopen/ooy012

33. Team RC, Team RC, Others. RA language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.-References. Published online 2020.

34. Pinheiro J, Bates D, R Core Team. nlme: Linear and Nonlinear Mixed Effects Models. Published online 2023. https://CRAN.R-project.org/package=nlme

35. Pinheiro JC, Bates DM, eds. Linear Mixed-Effects Models: Basic Concepts and Examples. In: *Mixed-Effects Models in S and S-PLUS*. Springer New York; 2000:3-56. doi:10.1007/0-387-22747-4_1

36. Data MC. *Secondary Analysis of Electronic Health Records*. Springer; 2016. https://play.google.com/store/books/details?id=qtlCDwAAQBAJ

37. Fawzy A, Ali H, Dziedzic PH, et al. Skin Pigmentation and Pulse Oximeter Accuracy in the Intensive Care Unit: a Pilot Prospective Study. *medRxiv*. Published online November 17, 2023. doi:10.1101/2023.11.16.23298645

38. Doshi T. Improving skin tone representation across. Google. Published May 11, 2022. Accessed February 8, 2024. https://blog.google/products/search/monk-skin-tone-scale/

39. Sjoding MW, Iwashyna TJ, Valley TS. Change the Framework for Pulse Oximeter Regulation to Ensure Clinicians Can Give Patients the Oxygen They Need. *Am J Respir Crit Care Med*. 2023;207(6):661-664. doi:10.1164/rccm.202209-1773ED

40. Dempsey K, Lindsay M, Tcheng JE, Ian Wong AK. The High Price of Equity in Pulse

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Oximetry: A cost evaluation and need for interim solutions. *medRxiv*. Published online September 23, 2023. doi:10.1101/2023.09.21.23295939

41. Vyas DA, Eisenstein LG, Jones DS. Hidden in Plain Sight - Reconsidering the Use of Race Correction in Clinical Algorithms. *N Engl J Med*. 2020;383(9):874-882. doi:10.1056/NEJMms2004740

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Tables

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Table 1. Characteristics of the cohort obtained after applying inclusion and exclusion criteria, grouped by race

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, Monk scale, Fitzpatrick scale, Von Luschan scale, and Delfin Melanin Index are ordered numerically ascending from light to dark, the other ones are ascending.

			Patients grouped by race group			
		Missing	Black	Other	White	Overall
n			57	15	56	128
Ethnicity, n (%)	Not Hispanic/Latino	0	57 (100.0)	12 (80.0)	52 (92.9)	121 (94.5)
	Hispanic/Latino			1 (6.7)	3 (5.4)	4 (3.1)
	Unknown			2 (13.3)	1 (1.8)	3 (2.3)
Gender, n (%)	Female	0	23 (40.4)	3 (20.0)	25 (44.6)	51 (39.8)
Observed oximeter location, n (%)	Forehead	0	1 (1.8)	0 (0.0)	0 (0.0)	1 (0.8)
	Missing		18 (31.6)	7 (46.7)	20 (35.7)	45 (35.2)
	Palm average		36 (63.2)	7 (46.7)	36 (64.3)	79 (61.7)
	Right toe		2 (3.5)	1 (6.7)	0 (0.0)	3 (2.3)
Fitzpatrick scale, mean (SD)		0	4.8 (0.6)	3.6 (1.1)	2.7 (0.6)	3.7 (1.2)
Von Luschan scale, mean (SD)		0	27.3 (2.3)	22.0 (5.2)	19.0 (3.4)	23.0 (5.1)
Monk scale, mean (SD)		0	6.4 (0.6)	5.2 (1.4)	4.3 (0.7)	5.3 (1.2)
Delfin ITA, mean (SD)		0	-2.9 (13.9)	21.0 (12.5)	37.2 (11.9)	17.4 (22.8)
Delfin L*, mean (SD)		0	48.9 (4.8)	56.4 (6.6)	63.1 (4.3)	56.0 (8.2)
Variable L*, mean (SD)		29	49.9 (5.1)	57.7 (5.5)	63.1 (4.0)	57.1 (7.8)
Konica Minolta L*, mean (SD)		8	40.4 (4.9)	48.2 (5.0)	54.4 (4.6)	47.5 (8.1)
Delfin Melanin Index, mean (SD)		0	742.2 (42.3)	650.2 (51.7)	598.4 (49.4)	668.5 (82.4)

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

Table 2. Results of the adjusted linear mixed-effects models

Results of the four linear mixed-effects models with clinical variables (SaO_2, pH heart rate, and MAP) adjusted, (Supplemental Formulas 5 to 9). Likelihood ratio tests (LRT) 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, □*2* statistics, and p-values are derived from LRT test results. N is the sample size of each model. Green cells represent negative coefficient values, i.e. the variable has an effect towards 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.

Expected Total Effect[1]: 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 in gray, above.

The self-reported race alone (Supplemental Formula 5) presents coefficients in the expected direction (-0.23%, 95% CI: -0.76%, 0.30%; p-value: 0.64 for Black patients, compared to White patients), but the p-value is not significant. When assessing the effect of a separate skin tone scale on bias (Supplemental Formula 6), only the Monk skin tone scale is shown to be significant (-2.40%; 95% CI: -4.32%, -0.48%; p-value: 0.01). The effect of all combined six skin tone scales on bias (the ones without missingness, Supplemental Formula 7) was found to be significant, with an expected total effect [1] of -1.72%, p-value of 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-value = 0.06).

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

Model	Variables	Coefficients	Lower 95% CI	Upper 95% CI	Chi-Squared	P-value	N
Association between race and bias (Supplemental Formula 5)	White	Baseline			0.90	0.64	463
	Black	-0.23	-0.76	0.30			
	Other	-0.31	-1.31	0.69			
Association between separate skin tone scale and bias (Supplemental Formula 6)	Fitzpatrick	-0.75	-2.07	0.57	1.37	0.24	463
	Von Luschan	-1.10	-2.68	0.48	1.99	0.16	463
	Monk	-2.40	-4.32	-0.48	6.07	**0.01**	463
	Delfin ITA	-0.62	-2.33	1.09	0.58	0.45	463
	Delfin L*	0.06	-1.70	1.82	<0.001	0.98	463
	Konica Minolta L*	-0.77	-2.58	1.04	0.76	0.38	424
	Variable L*	-0.31	-2.03	1.41	0.18	0.67	367
	Delfin Melanin Index	0.17	-1.53	1.87	0.03	0.87	463
Association between six of skin tone scales and bias (Supplemental Formula 7)	Fitzpatrick	0.37	-2.27	3.01	14.98	**0.02**	463
	Von Luschan	0.36	-2.53	3.25			
	Monk	-3.55	-6.93	-0.17			
	Delfin ITA	-6.62	-12.22	-1.02			
	Delfin L*	4.66	-0.40	9.72			
	Delfin Melanin Index	3.06	-1.15	7.27			
	Expected Total Effect[1]	**-1.72**	-1.72	-1.72			
Association between eight of skin tone scales and bias (Supplemental Formula 8)	Fitzpatrick	-0.76	-4.00	2.48	14.86	0.06	328
	Von Luschan	2.53	-1.03	6.09			
	Monk	-3.43	-7.61	0.75			
	Delfin ITA	-7.61	-14.40	-0.82			
	Delfin L*	3.58	-2.06	9.22			
	Konica Minolta L*	-1.92	-5.58	1.74			

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

	Variable L*	0.75	-3.46	4.96			
	Delfin Melanin Index	3.06	-1.82	7.94			
	Expected Total Effect[1]	**-3.80**	-3.80	-3.80			

Figures

Figure 1. Flow diagram

A total of 1,167 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 out 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 didn't have complete skin measurement data, and four patients didn't 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 didn't have pulse oximetry locations specified, we presumed the measurement was done on the finger and imputed it using the four palm locations as well.

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .

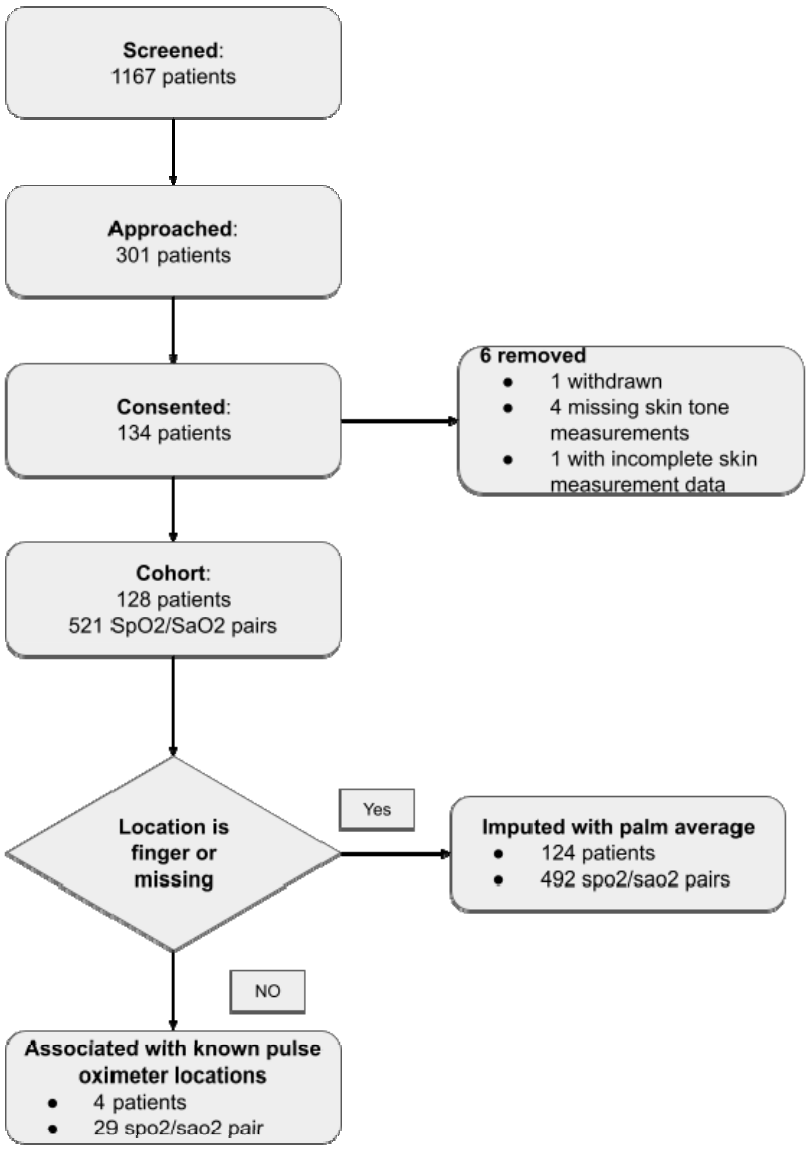
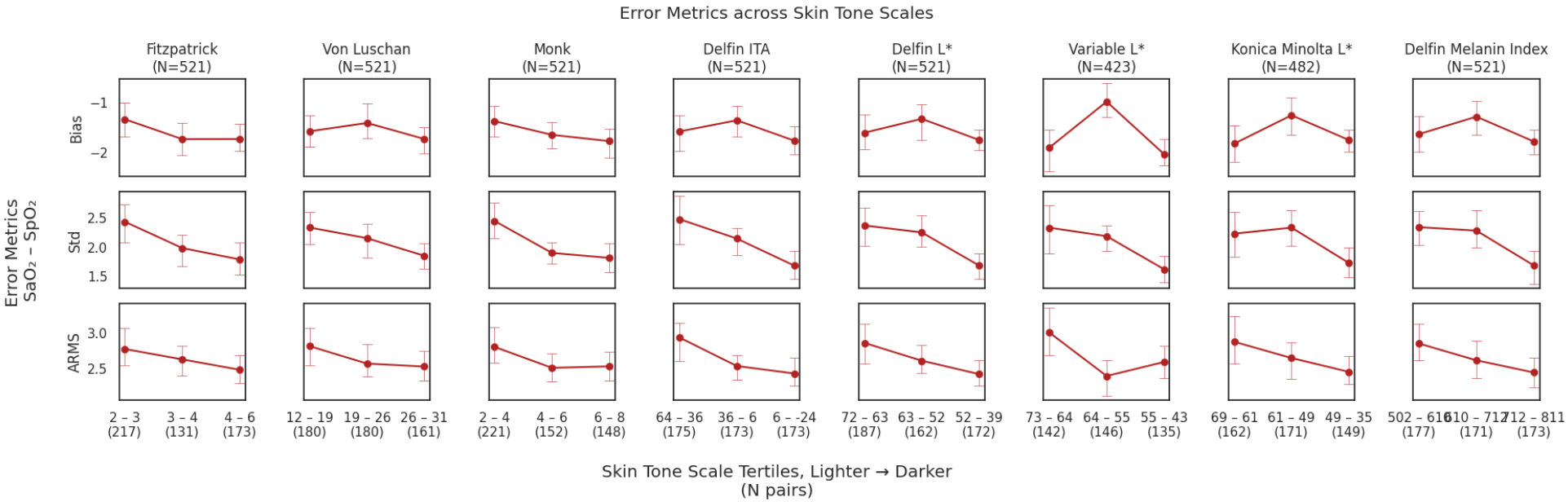


Figure 2. Unadjusted error metrics in pulse oximetry across skin tone scale tertiles

Unadjusted error metrics of mean directional bias, standard deviation, and accuracy root mean square (also known as A_{RMS} or root mean square error), across tertiles. Tertiles are ordered from lightest to darkest, from the left to the right on the x-axis. Note that a pulse oximetry bias defined as SaO_2 - SpO_2 results in a negative bias reflecting that pulse oximetry overestimates true oxygenation values. Fitzpatrick and Monk appear to have a trend towards more negative bias (e.g., bias increasingly negative) from lighter to darker tertiles. A_{RMS} appears to be lower (that is, a lower root mean square error) in many darker tertiles than in lighter tertiles. Variable L* and Konica Minolta L* have fewer patients because there was more missingness. Some patients did not have these measurements either due to patient refusal (often due to feelings of being overwhelmed, stress, or experiencing pain) or interruptions by clinical workflow.



medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license .

medRxiv preprint doi: https://doi.org/10.1101/2024.02.24.24303291; this version posted February 27, 2024. The copyright holder for this preprint
(which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
It is made available under a CC-BY-NC-ND 4.0 International license .