Measurement Bias in Wearable PPG Sensors:

Ethical and Policy Implications for AI-Enabled Remote Patient Monitoring

Akintoye Oyedola
School of Computing, University of North Florida

CNT6167 Internet of Things; Literature Review

Abstract—In 2025, one in five dark-skinned patients in an EquiOx study had undetected hypoxemia because the pulse oximeter was built to fail them. This review examines how that hardware failure propagates through AI-enabled remote patient monitoring (RPM) pipelines and why, four years after the FDA acknowledged the problem, no binding pre-market standard exists. Wearable photoplethysmography (PPG) sensors carry a physics-based measurement bias rooted in melanin absorption of light at 660 nm and 940 nm. AI models trained on this biased data inherit and amplify the error at every pipeline stage, ultimately withholding supplemental oxygen from hypoxic patients. Five critical gaps remain open: no standardized fairness benchmark, no finalized validation protocol, limited research-to-deployment evidence, weak informed consent frameworks, and unrecognized socioeconomic compounding. The central argument is that algorithmic sophistication cannot compensate for signal that was never accurately captured.

Keywords—photoplethysmography, PPG bias, pulse oximetry, skin tone, remote patient monitoring, algorithmic fairness, health equity, FDA regulation, EU AI Act

1. Introduction

Wearable PPG sensors measure heart rate, peripheral oxygen saturation (SpO_2), and Heart Rate Variability (HRV), feeding clinical decision support models for disease prediction, risk stratification, and treatment recommendations [1]. However, a documented hardware limitation undermines the equity of these systems. Melanin absorbs LED light at the wavelengths pulse oximeters use (660 nm and 940 nm), producing systematically less accurate SpO_2 readings for darker-skinned patients [2]. During COVID-19, Sjoding et al. [3] found Black patients experienced nearly three times the rate of occult hypoxemia compared to White patients. Occult Hypoxemiaa is a critical medical discrepancy where a patient's actual arterial blood oxygen level is low (<88%) while a pulse oximeter incorrectly displays a normal reading (92% – 98%) [4]. The 2025 EquiOx study confirms this bias persists in modern devices [5].

When PPG data from non-representative populations trains AI risk models, those models inherit the hardware error. Obermeyer et al. [6] showed a widely deployed algorithm required Black patients to be significantly sicker than White patients to receive the same care recommendation. This is a direct consequence of biased inputs. The FDA issued a non-binding Safety Communication in 2021 [7] and floated pre-market testing in a 2023 discussion paper with no finalized solutions. The EU AI Act [8] imposes enforceable data-governance obligations on high-risk AI but lacks a PPG-specific hardware standard. The WHO [9] provides ethical guidance but lacks enforcement power.

This survey traces the full pipeline from sensor physics to policy gap around five research questions:

- RQ1: how PPG sensor characteristics introduce systematic bias
- RQ2: how AI practices amplify that bias across demographic groups
- RQ3: what fairness interventions have been proposed
- RQ4: how bias manifests as allocative harm and consent gaps
- RQ5: whether current regulatory frameworks can require a hardware-level fix

This literature review study argues that the answer to RQ5 is no. This measurement bias issue became urgent to me when I found that the patients most harmed are also the least likely to appear in studies that could change it.

2. Related Work and Literature Categorization

2.1. Group 1: Clinical and Empirical Evidence of PPG Bias

Coppetti et al. [2] establish the optical physics: melanin reduces signal-to-noise at 660 nm and 940 nm, causing systematic SpO_2 overestimation. This is a hardware design problem. Sjoding et al. [3] quantified the clinical consequence at scale: a threefold occult hypoxemia rate in Black patients. The 2025 EquiOx study [5] confirms this on contemporary hardware (20% error rate), and AlSaafeen et al. [10] extend the finding to consumer smartwatches, confirming the problem spans both clinical-grade and consumer-grade devices.

2.2. Group 2: Technical and AI Modeling Literature

Majumder et al. [11] establish that LED wavelength selection is a structural design constraint as PPG sensor wavelengths were never chosen with skin tone diversity in mind. Böttcher et al. [12] prove that motion artifact filtering in wearables can discard up to 50% of data, disproportionately excluding signals from higher-movement populations and introducing sampling bias before training begins. Nazir et al. [1] map the full IoT-to-clinical-AI pipeline and note that black-box deep learning architectures make it nearly impossible to trace a prediction error back to the sensor input that caused it. Mehrabi et al. [13] provide the bias taxonomy: *measurement bias* (hardware signal degradation from melanin) and *representation bias* (underrepresentation in training data) require different fixes. Obermeyer et al. [6] demonstrate the downstream harm: biased input data forced a deployed algorithm to require Black patients to appear significantly sicker before receiving equivalent care.

2.3. Group 3: Policy, Ethics, and Industry Frameworks

The FDA's 2021 Safety Communication [7] explicitly acknowledges SpO_2 inaccuracy by skin tone but issued only non-binding guidance. The 2023 discussion paper proposed Fitzpatrick-scale testing but never mandated it. The FDA AI/ML SaMD Action Plan [14] introduces Good Machine Learning Practice (GMLP) for algorithmic bias broadly, but does not impose device-specific sensor validation requirements. The EU AI Act [8] is binding (penalties up to €30M or 6% of global turnover) and requires bias-testing documentation. However, it does not provide a PPG-specific optical sensor standard. The WHO [9] defines non-discrimination and transparency in AI, but does not have the authority to enforce any policies. Intel's 2022 white paper [15] represents current industry best practice (federated learning, synthetic data), which this paper argues addresses the symptom rather than the source: no model-level fix can recover a signal that the hardware never captured accurately.

3. Methodology

The sources were retrieved from IEEE Xplore, ACM Digital Library, Scopus, UNF OneSearch, PubMed, and government/industry portals (FDA.gov, EUR-Lex, WHO). Seventeen sources were selected, covering peer-reviewed articles, regulatory communications, conference proceedings, and an industry white paper. Inclusion required relevance to at least one of four pipeline stages: PPG sensor hardware, data pre-processing, AI modeling, or policy/ethics governance. Table 1 shows the full criteria. The selected sources were analyzed on five dimensions: IoT data ecosystem, AI models, bias and fairness, ethics, and policy. The organizing argument, that hardware bias propagates through the AI pipeline and that current regulation cannot require a hardware-level fix, was tested against each source.

4. Results and Analysis

4.1. Bias Propagation Pipeline

Table 2 maps each pipeline stage to its type of bias and patient impact. Hardware measurement bias at the sensor stage compounds at every downstream stage, arriving at allocative harm, i.e., withheld oxygen, in the clinical outcome. The core finding is that the measurement bias error is not introduced by the AI; it is baked in before the first data point is collected.

4.2. Regulatory Evaluation

Table 3 compares the three frameworks on the specific question of PPG sensor skin tone validation. The FDA's Safety Communication is unique in U.S. regulation because it names skin tone explicitly, yet has

Table 1: **Inclusion and Exclusion Criteria**

Inclusion Criteria	
Device Scope	PPG-based wearables, pulse oximeters, or smartwatches in clinical or RPM contexts
AI/Signal Model	ML models or signal processing applied to PPG-derived data (SpO_2, HR, HRV)
Bias Focus	Bias or accuracy disparities by skin tone (Fitzpatrick scale), race, or body composition
Ethics/Policy	Allocative harm, regulatory frameworks (FDA, EU AI Act), or data protection in PPG/RPM contexts
Exclusion Criteria	
Non-PPG Primary	Studies focused on ECG, accelerometers, or EMG without PPG as primary device
Non-Clinical	Consumer devices without health monitoring claims or clinical validation
No Bias Analysis	Hardware optimization without demographic bias analysis
Generic AI	Theoretical bias not applied to physiological sensor data

Table 2: **PPG Pipeline: Bias Type by Stage**

Stage	**Practice**	**Bias Type**	**Patient Impact**
Sensor Design	LED uncalibrated across Fitzpatrick scale	Measurement	Inaccurate SpO_2; missed hypoxemia [2,3]
Pre-processing	Motion filtering drops PPG segments	Sampling	Reduced usable data for high-movement groups [12]
Labeling	SpO_2 ground truth from biased sensor	Label	AI inherits overestimation; flags hypoxia as normal [3]
Training	Global loss without subgroup constraints	Algorithmic	Subgroup accuracy sacrificed [13]
Deployment	Risk scores from biased SpO_2	Allocative Harm	Oxygen withheld; 20% error rate [5]

produced no binding pre-market requirement in over four years. The EU AI Act is a genuine improvement as its financial penalties create structural incentives for bias testing. Yet, a manufacturer can satisfy its documentation requirements while still shipping a sensor that overestimates SpO_2 for darker-skinned patients. No existing framework closes this gap.

4.3. Key Trends and Identified Gaps

Three trends emerge from the literature. First, hardware-level SpO_2 overestimation in darker-skinned patients is consistent from Coppetti et al.'s 2019 physics characterization [2] to the 2025 EquiOx prospective result [5]. Incremental hardware updates have not fixed the fundamental calibration problem. Second, AI pipelines amplify rather than correct this error due to biased labels corrupt training, and black-box architectures that obscure where the error originated [1,13]. Consequently, the harm materializes as oxygen withheld from patients the model never flagged as hypoxic [6]. Finally, regulators have acknowledged the problem without resolving it: the FDA since 2021, the EU Act still without a PPG-specific standard [8,14].

Five gaps remain open. (1) No annotated PPG dataset with Fitzpatrick skin tone scores exists, making cross-study verification of bias mitigation impossible. (2) The FDA's 2023 discussion paper named the need for diverse testing but specified no sample sizes, thresholds, or protocols. (3) Fairness interventions have been evaluated almost exclusively in controlled settings with little real-world clinical evidence [16]. (4) Patients whose PPG data trains AI models are rarely informed of this use; consent frameworks vary widely by institution and jurisdiction [17]. (5) Patients most affected by sensor bias are also more likely to use

Table 3: **Regulatory Comparison: PPG Skin Tone Validation**

Feature	**FDA AI/ML Plan**	**FDA Pulse Ox Safety Comm.**	**EU AI Act (2024)**
Legal Status	Non-binding	Safety Comm. + non-binding paper	Binding
Skin Tone	No PPG language	Explicitly names SpO_2 inaccuracy	No PPG hardware standard
Bias Testing	GMLP; no enforcement	Proposes Fitzpatrick testing; no mandate	Requires documentation
Penalty	None	None	€30M or 6% turnover
Key Gap	Treats as general AI	Acknowledges; no mandate	No optical sensor standard

lower-grade consumer devices [18]. This is a compounding dynamic absent from both the technical and regulatory literature.

5. Discussion

The FDA's acknowledgement of the PPG sensor measurement bias is the most striking regulatory finding. No other U.S. document for a specific medical device names skin tone bias as explicitly as the 2021 Safety Communication [7]. Yet, that acknowledgment has not produced any binding pre-market requirement in four years. While the EU AI Act [8] implements penalties, its scope stops at model governance without specific PPG sensor mandates. Thus, a manufacturer can satisfy every EU documentation requirement while deploying a sensor whose optical calibration was never validated across skin tones. That is the gap neither framework closes.

The Intel white paper [15] represents current industry best practice, including federated learning and synthetic data generation. These are useful at the model stage, but they do not fix hardware. Federated learning diversifies training data, but it does not correct the measurement error before it enters those data. Synthetic augmentation cannot reproduce the physiological signal characteristics of darker skin if the underlying problem is optical and physics-based. This is consistent with Böttcher et al. [12] on the loss of data quality of hardware-origin and Majumder et al. [11] on the wavelength of LED as a structural constraint. No amount of downstream sophistication can recover a signal that the sensor never captured.

6. Recommendations

Five targeted recommendations follow from the analysis. **R1:** The FDA should finalize binding pre-market accuracy requirements for PPG devices across all Fitzpatrick categories, using arterial blood gas, not oximeter self-comparison, as ground truth. **R2:** A public Fitzpatrick-annotated PPG benchmark dataset must be developed under NIH or equivalent sponsorship. Without it, bias reduction claims cannot be independently verified. **R3:** FDA and EU regulatory submissions for AI-enabled RPM tools should require disaggregated performance reporting by skin tone, applied to both the model and the sensor that feeds it. **R4:** The EU AI Act's high-risk medical AI classification should explicitly cover the sensor hardware powering those systems, making optical calibration enforceable alongside model governance. **R5:** Institutions deploying RPM systems that feed patient PPG data into AI pipelines should implement explicit consent protocols with an opt-out that does not affect clinical care [17].

7. Conclusion

This literature review traced the causal chain of measurement bias from PPG sensor optics through AI training pipelines to clinical outcomes and regulatory governance. The bias is not correctable by algorithmic change alone because it is a signal that was never accurately captured. The FDA acknowledged it in 2021 and has not mandated a fix. While the EU AI Act is stricter, it stops at model documentation, leaving the hardware gap open. Until a binding Fitzpatrick-scale validation standard exists, and until a public benchmark enables independent verification of mitigation claims, marginalized patients will continue to be monitored by devices that were never tested on them. Ultimately, the technical path forward is clear, and what is needed is the regulatory will to require it.

References

[1] S. Nazir, Y. Ali, N. Ullah, and I. García-Magariño, "Internet of things for healthcare using artificial intelligence: Opportunities and challenges," *IEEE Access*, vol. 7, pp. 128516–128541, 2019.

[2] R. Coppetti, A. Braghieri, S. Brambilla, and A. Calvo, "Correction of SpO_2 values measured with a pulse oximeter in patients with dark skin pigmentation," *Journal of Clinical Monitoring and Computing*, vol. 33, pp. 609–616, 2019.

[3] M. W. Sjoding, R. P. Dickson, T. J. Iwashyna, S. E. Gay, and T. S. Valley, "Racial bias in pulse oximetry measurement," *New England Journal of Medicine*, vol. 383, pp. 2477–2478, Dec. 2020.

[4] N. J. Parr, "Differential pulse oximeter accuracy, occult hypoxemia prevalence, and clinical outcomes by patient race/ethnicity: A systematic review."

[5] C. Hendrickson *et al.*, "Pulse oximeter performance varies by skin pigment: Results from the EquiOx study," in *American College of Cardiology Scientific Session*, (Chicago, IL), Mar. 2025.

[6] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[7] U. Food and D. Administration, "Pulse oximeters: Recommendations for patients, health care providers, and developers." FDA Safety Communication, Feb. 2021. [Online]. Available: `https://www.fda.gov/medical-devices/safety-communications/pulse-oximeters-recommendations-patients-health-care-providers-and-developers`.

[8] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence," *Official Journal of the European Union*, vol. L, 2024. Text with EEA relevance.

[9] World Health Organization, *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization, Geneva, Switzerland, 2021.

[10] S. Asif, A. AlSaafeen, S. Nadar, S. Nambiar, J. Dannawi, H. Korrapati, and H. S. Wilkhoo, "Photoplethysmography in diverse skin tones: Evaluating bias in smartwatch health monitoring," *Cureus*, vol. 17, Oct 2025.

[11] S. Majumder, T. Mondal, and M. J. Deen, "Wearable sensors for remote health monitoring," *Sensors*, vol. 17, no. 1, p. 130, 2017.

[12] S. Böttcher, S. Vieluf, J. Bruno, *et al.*, "Data quality of multimodal wearable monitoring: A large-scale field study," *IEEE Internet of Things Journal*, vol. 9, no. 24, pp. 25110–25122, 2022.

[13] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021.

[14] U.S. Food and Drug Administration, "Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan," white paper, U.S. Food and Drug Administration, Washington, DC, Jan. 2021.

[15] Intel Health and Life Sciences, "Addressing bias in AI for healthcare: Technical and ethical considerations," white paper, Intel Corporation, 2022.

[16] B. A. Mir, S. R. Abbas, and S. W. Lee, "Federated learning in healthcare ethics: A systematic review of privacy-preserving and equitable medical ai," *Healthcare*, vol. 14, p. 306, Jan 2026.

[17] S. Yu, S.-S. Lee, and H. Hwang, "The ethics of using artificial intelligence in medical research," *Kosin Medical Journal*, vol. 39, p. 229–237, Dec 2024.

[18] J. Joseph, "Algorithmic bias in public health ai: A silent threat to equity in low-resource settings," *Frontiers in Public Health*, vol. 13, Jul 2025.