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
Jacksonville, Florida

***Abstract*—In 2025, one in five dark-skinned patients in an EquiOx study had undetected hypoxemia directly because the pulse oximeter sensor was built to fail them. This literature review examines how that hardware failure propagates through AI-enabled remote patient monitoring pipelines and why, four years after the FDA acknowledged the problem, no binding pre-market standard exists.**

Wearable photoplethysmography (PPG) sensors carry a documented, physics-based measurement bias rooted in the optical absorption properties of melanin at 660 nm and 940 nm. When AI models for risk stratification and clinical decision support are trained on data from these sensors, they inherit and amplify this error at every pipeline stage. This manifests in corrupted training labels through black-box deep learning architectures to clinical outputs that withhold supplemental oxygen from hypoxic patients.

Regulatory responses have acknowledged the problem without resolving it. The FDA issued a non-binding Safety Communication in 2021 but never finalized a pre-market mandate. The EU AI Act imposes binding data-governance obligations but contains no PPG-specific optical sensor standard. This paper identifies five critical gaps: no standardized fairness benchmark, no finalized validation protocol, limited research-to-deployment evidence, weak informed consent frameworks, and unrecognized socioeconomic compounding. This paper also proposes five targeted recommendations. The central argument is that algorithmic sophistication cannot compensate for signal that was never accurately captured. Every day this gap remains open, and marginalized patients are being monitored by devices their designers never tested on them.

***Index Terms*—photoplethysmography, PPG bias, pulse oximetry, skin tone, remote patient monitoring, wearable sensors, algorithmic fairness, health equity, FDA regulation, EU AI Act**

I. Introduction

Wearable photoplethysmography (PPG) sensors are a primary data source for AI-enabled remote patient monitoring (RPM) systems. These devices measure heart rate, blood oxygen saturation (SpO_2), and heart rate variability (HRV), providing longitudinal physiological data streams that feed clinical decision support models for disease prediction, risk stratification, and treatment recommendations [1]. The deployment of these systems in hospital settings and consumer wearables has positioned PPG-derived data as a key input to a new generation of AI in healthcare.

However, a well-documented hardware limitation undermines the equity of these systems. The optical physics underlying the PPG measurement produces systematically less accurate SpO_2 readings for patients with a higher melanin concentration [2]. Specifically, LED emission is absorbed and reflected by tissue at different characteristic wavelengths. This is a structural bias encoded in the device design before patient data is collected. The clinical consequences were documented at scale during the COVID-19 pandemic, when Sjoding et al. [3] found that Black patients experienced nearly three times the rate of occult hypoxemia compared to White patients. Occult hypoxemia is defined as low arterial oxygen saturation undetected by pulse oximeters. The 2025 EquiOx prospective study confirms that this bias persists in modern devices [4].

When PPG data from non-representative populations is used to train AI models for risk stratification or clinical decision support, the AI models inherit and amplify this hardware error. Obermeyer et al. [5] demonstrated that a widely deployed risk-stratification algorithm effectively requires Black patients to be significantly sicker than White patients to receive the same care recommendation. This phenomenon is a direct consequence of biased input data. The pattern of allocative harm, where sensor failure translates into withheld treatment, is the central problem this paper addresses.

The regulatory response has been inadequate. While the FDA issued a non-binding Safety Communication on pulse oximetry skin tone bias in 2021 [6] and proposed pre-market testing requirements in a 2023 discussion paper, it did not finalize any policies. The EU AI Act imposes enforceable data-governance obligations on high-risk AI systems [7] but does not contain a hardware validation standard specific to PPG. The WHO provides ethical guidance on non-discrimination and accountability for AI in healthcare [8], but lacks enforcement power.

This paper traces that failure from the LED to the policy gap. The analysis begins with the optical physics that make PPG sensors structurally less accurate for darker-skinned patients, moves through the AI modeling practices that compound this hardware error at every pipeline stage, and arrives at a comparative evaluation of the regulatory frameworks that have acknowledged the problem for years without resolving it. The main research question is not whether bias exists but whether current governance frameworks are structurally

capable of requiring a hardware-level validation policy to fix it. The answer, this paper argues, is no, and the reforms needed are specific and achievable. This question became urgent to me when I discovered that the patients most affected by this bias are also the least likely to be included in studies that could fix it.

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

- **RQ1 (Data):** How do PPG sensor characteristics (LED wavelength and melanin absorption) introduce systematic measurement bias, and how does this affect data quality for downstream AI models?
- **RQ2 (Models):** How do AI modeling practices interact with sensor-level bias to amplify disparities in model performance across demographic groups?
- **RQ3 (Fairness):** What fairness metrics and technical interventions have been proposed or evaluated for PPG-derived healthcare AI?
- **RQ4 (Ethics):** How does PPG measurement bias manifest as allocative harm, transparency failures, and informed consent gaps?
- **RQ5 (Policy):** Do existing regulatory frameworks require pre-market sensor-level validation across diverse skin tones, and what reforms would close the gap?

II. Related Work and Literature Categorization

The literature is organized into three functional groups that together trace the data-to-policy pipeline from sensor hardware to governance.

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

These studies establish that PPG measurement bias is real and clinically consequential, not merely theoretical.

Coppetti et al. [2] characterize the optical physics underlying the bias. Melanin in the dermis absorbs light at 660 nm (red) and 940 nm (near-infrared), reducing the signal-to-noise ratio for patients with higher melanin concentration and causing systematic SpO_2 overestimation. 660 nm (red) and 940 nm (near-infrared) are the wavelengths used by pulse oximeters. This is a hardware design problem that establishes that sensor calibration and wavelength selection must be part of any comprehensive fairness intervention.

Sjoding et al. [3] provide the landmark clinical evidence from a large hospital dataset collected during the COVID-19 pandemic. Black patients had nearly three times the rate of occult hypoxemia compared to White patients, directly translating hardware measurement error into withheld treatment at scale. The 2025 EquiOx study [4] confirms this disparity persists on contemporary hardware, reporting an error rate of 20% occult hypoxemia in dark-skinned patients. AlSaafeen et. al [9] extend this finding to consumer smartwatches, confirming that the problem spans both clinical-grade and consumer-grade PPG devices.

B. Group 2: Technical and AI Modeling Literature

Majumder et al. [10] provide a systematic analysis of physical constraints of the wearable sensor, defining the selection of LED wavelength as a design constraint rather than an anomaly. The critical observation is that wavelength choices were not made with skin tone diversity as a design requirement. Therefore, PPG measurement bias is structural; built into devices before any patient interaction.

Böttcher et al. [11] demonstrate in a large-scale field study that data loss in wearable devices can reach 50% and is strongly influenced by motion artifact filtering. In the context of PPG, aggressive filtering disproportionately excludes signal windows from patients with higher baseline movement variability, introducing sampling bias at the pre-processing stage before AI training begins.

Nazir et al. [1] provide a comprehensive architectural blueprint of the pipeline from raw IoT sensors to clinical AI models. This blueprint identifies where fairness interventions are technically feasible. Notably, deep learning models applied to PPG time-series data are typically black-box in nature, making it difficult to trace a prediction error back to the specific sensor input that caused it. This attribute compounds the harm from measurement bias.

Mehrabi et al. [12] provide the main conceptual tool to categorize bias types. Of more than 20 definitions, two map directly into the PPG use case: *measurement bias*, which is hardware-level signal degradation caused by melanin absorption, and *representation bias*, the under-representation of darker-skinned patients in PPG training datasets. Both require different remediation strategies, making precise taxonomy important.

Obermeyer et al. [5] demonstrate the downstream consequence of biased IoT inputs. In a widely deployed risk-stratification algorithm, healthcare spending was used as a proxy for illness, but because Black patients received less care at a given level of illness (a product of systemic inequity), the algorithm required them to be significantly sicker than White patients to receive the same care recommendation.

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

The FDA's 2021 Safety Communication [6] is the primary U.S. regulatory evidence of acknowledged skin tone bias without resolution. The FDA explicitly stated that pulse oximeters may be less accurate for people with darker skin tones and issued non-binding recommendations to patients and clinicians. A 2023 FDA discussion paper proposed potential Fitzpatrick-scale testing requirements but did not reach a final mandate.

The FDA AI/ML SaMD Action Plan [13] addresses algorithmic bias broadly across all software as a medical device. It introduces the concept of Good Machine Learning Practice (GMLP). It also acknowledges the need to address bias in AI algorithms, but does not issue device-specific mandates for sensor hardware validation across demographic groups.

The EU AI Act [7] imposes binding data-governance obligations on high-risk AI systems, including medical diagnostics,

with penalties up to €30 million or 6% of global turnover for non-compliance. It requires technical documentation of bias testing prior to market access, but does not contain a PPG-specific optical sensor standard.

The WHO guidance [8] defines the ethical standard for AI in healthcare with the core principles of non-discrimination, transparency, and accountability. It provides the ethical vocabulary against which both FDA and EU frameworks can be assessed. However, the WHO guidelines do not carry enforcement power.

Intel's 2022 industry white paper [14] describes federated learning and synthetic data generation as practical tools to reduce bias in clinical AI. It represents the current best-practice industry position, which this paper argues is insufficient when the bias is hardware-level: no algorithmic sophistication can fully recover signal that was never accurately captured.

III. Methodology

A. Literature Search Strategy

The sources were retrieved from IEEE Xplore, ACM Digital Library, Scopus, UNF OneSearch, PubMed, and direct government and industry portals (FDA.gov, EUR-Lex, WHO publications). Twelve core artifacts were selected: peer-reviewed academic articles, conference proceedings, regulatory communications, and an industry white paper. Inclusion required relevance to at least one of the four pipeline stages: PPG sensor hardware, data collection and pre-processing, AI modeling, and policy or ethics governance.

Four keyword search blocks:

1) PPG & wearable sensors: "photoplethysmography," "pulse oximeter," "SpO2 accuracy," "remote patient monitoring"
2) Skin tone & measurement bias: "melanin," "Fitzpatrick scale," "skin tone bias," "LED wavelength"
3) AI bias & fairness: "measurement bias," "algorithmic bias," "health equity," "demographic disparities"
4) Regulation & ethics: "FDA pulse oximeter," "EU AI Act medical device," "Good Machine Learning Practice," "allocative harm"

B. Inclusion and Exclusion Criteria

Table I summarizes the inclusion and exclusion criteria applied to candidate sources.

C. Analytical Framework

I analyzed the twelve sources along a five-dimensional taxonomy:

1) *IoT Data Ecosystem*: device types, data characteristics, and structural sources of bias, including digital divides and sensor performance variation.
2) *AI Models*: learning algorithms, pre-processing pipelines, and deployment settings.
3) *Bias and Fairness*: types of bias, empirical disparities, and mitigation methods.
4) *Ethics*: fairness, transparency, consent, and accountability.

TABLE I
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 pipelines applied to PPG-derived data (SpO_2, HR, HRV)
Bias Focus	Explicit evaluation of 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 primarily on ECG, accelerometers, or EMG without PPG as a primary device
Non-Clinical	Consumer devices without health monitoring claims or clinical validation
No Bias Analysis	Hardware optimization studies without demographic bias analysis
Generic AI	Theoretical bias discussions not applied to physiological sensor data

5) *Policy*: data protection, medical device regulation, and emerging AI governance.

The organizing argument is that wearable PPG sensors have a documented physics-based measurement bias that disproportionately harms patients with darker skin tones. This hardware-level bias propagates through AI training pipelines into clinical decision support systems. Furthermore, current regulatory frameworks, despite acknowledging the problem, have not issued binding mandates for sensor-level demographic validation prior to market access.

IV. Results and Analysis

A. Bias Propagation Pipeline

Table II maps each stage of the PPG data pipeline to its associated bias type and patient impact. The central finding is that hardware measurement bias at the sensor stage propagates and compounds at every subsequent stage, ultimately producing allocative harm, i.e., withheld oxygen therapy, in the clinical outcome.

B. Classifier and Regulatory Evaluation

Table III presents a comparative evaluation of the three main regulatory frameworks applied to the specific question of PPG sensor skin tone validation.

The FDA's device-specific acknowledgment in the Pulse Oximetry Safety Communication is unique in the U.S. regulatory landscape as it names skin tone explicitly, unlike the broader Action Plan. Yet, it has produced no binding premarket requirement in more than four years. The enforceability of the EU AI Act is a genuine improvement over the FDA approach. However, its absence of a PPG-specific optical sensor standard means that manufacturers can technically comply with data-governance documentation requirements while still deploying sensors that overestimate SpO_2 in darker-skinned patients.

TABLE II
PPG TECHNICAL PIPELINE: FROM SENSOR DESIGN TO BIAS TYPE

Stage	Technical Practice	Bias Type	Patient Impact
Sensor Design	LED not calibrated across Fitzpatrick scale	Measurement Bias	Darker-skinned patients receive inaccurate SpO_2; missed hypoxemia [2], [3]
Signal Pre-processing	Motion artifact filtering excludes PPG signal segments	Sampling Bias	Systematically reduced usable data for high-movement populations [11]
Labeling	SpO_2 ground truth derived from biased sensor	Label Bias	AI inherits hardware overestimation; predicts normal oxygenation for hypoxic patients [3]
Model Training	Global loss minimization without skin-tone subgroup constraints	Algorithmic Bias	Model maximizes overall accuracy; sacrifices subgroup performance [12]
Deployment	Risk stratification based on biased SpO_2 inputs	Allocative Harm	Supplemental oxygen withheld; 20% error rate in prospective study [4]

TABLE III
REGULATORY FRAMEWORK COMPARISON: PPG SENSOR VALIDATION

Feature	FDA AI/ML Action Plan	FDA Pulse Oximetry Safety Comm.	EU AI Act (2024)
Legal Status	Non-binding guidance	Safety Comm. + non-binding discussion paper	Binding regulation
Skin Tone Language	No PPG-specific language	Explicitly acknowledges SpO_2 inaccuracy by skin tone	No PPG-specific hardware standard
Bias Testing	Recommends GMLP; no enforcement	Proposes (not mandates) Fitzpatrick-scale testing	Requires bias testing documentation
Penalty	None pre-market	None pre-market	Up to €30M or 6% global turnover
Key Gap	Treats PPG bias as general AI problem	Acknowledges for years; no binding mandate issued	Stricter than FDA but no optical sensor standard

C. Key Trends

Trend 1: Hardware-Level Bias Persists Over Time. Hardware-level SpO_2 overestimation in darker-skinned patients is consistent through a decade of PPG literature, from sensor physics characterizations in 2019 [2] through the 2025 EquiOx prospective study [4]. Clearly, incremental hardware improvements have not addressed the fundamental optical calibration problem. The 2025 EquiOx study's 20% occult hypoxemia error rate on contemporary devices indicates that this is not a historical artifact.

Trend 2: AI Pipelines Amplify Rather Than Correct Bias. AI pipelines trained on non-representative datasets inherit and amplify sensor error at every stage. Biased oximeter readings produce corrupted training labels and black-box deep learning architectures obscure the demographic source of prediction errors [1], [12]. Obermeyer et al. [5] demonstrate this pattern in an adjacent risk-stratification system. The harm manifests itself as allocative harm in which supplemental oxygen is withheld from hypoxic patients because neither the sensor nor the model flagged a problem.

Trend 3: Regulatory Acknowledgment Without Enforcement. Regulatory responses have acknowledged the PPG sensor bias problem without resolving it. The FDA explicitly named skin tone bias in 2021 yet has not issued a binding pre-market validation standard [13]. The EU AI Act imposes enforceable data-governance obligations but does not contain a PPG-specific optical sensor requirement [7].

D. Identified Gaps

No Standardized PPG Fairness Benchmark. Unlike algorithmic fairness domains with established public datasets, no annotated PPG dataset with Fitzpatrick skin tone scores exists. This makes independent verification and cross-study comparison of bias mitigation techniques currently impossible.

Missing Pre-Market Validation Protocol. The FDA's 2023 discussion paper acknowledged the need for skin-tone-diverse testing, but did not provide fundamental protocol guidelines about required sample sizes per Fitzpatrick category, testing conditions, and clinical-use-case-specific performance thresholds.

Research-to-Deployment Gap. Fairness interventions like federated learning and post-hoc calibration have been evaluated almost exclusively in controlled settings, with little published evidence on real-world performance in live clinical workflows [15].

Informed Consent and Data Ownership. Patients whose wearable PPG data trains clinical AI models are rarely informed of this use. The consent frameworks that govern the practice vary widely between institutions and jurisdictions without a PPG-specific standard [16].

Socioeconomic Compounding. The patients most affected by the PPG measurement error are also the most likely to be monitored with lower-grade consumer devices [17]. This compounding dynamic, where darker-skinned, lower-resource patients experience hardware bias and reduced device quality, is absent from both the technical literature and current regulatory frameworks.

V. DISCUSSION

A. Comparison to Prior Regulatory Frameworks

The FDA's position is internally inconsistent in a practical way. Its device-specific Safety Communication [6] acknowledges skin tone bias more explicitly than any other U.S. regulatory document for a specific medical device. Yet, this specificity has not translated into binding pre-market requirements. The broader AI/ML Action Plan [13] recommends GMLP principles that implicitly address demographic bias but

do not require manufacturers to validate the precision of the PPG on the Fitzpatrick scale before market access.

The EU AI Act [7] offers a genuinely stronger compliance architecture. Its binding requirements, enforceability, and financial penalties create structural incentives for manufacturers to invest in bias testing that the FDA's guidance model does not. However, compliance with the EU AI Act's data governance documentation does not require a manufacturer to demonstrate that their optical sensor is accurate across diverse skin tones; only that the AI model trained on that sensor's output has been documented. This is a meaningful gap: a model can be well-documented and compliant while being trained on systematically biased hardware data.

B. Technical Limitations and Remediation

The Intel white paper [14] represents the current best-practice industry position on algorithmic bias mitigation, describing federated learning and synthetic data generation as practical tools. This paper argues that these approaches address the symptom rather than the cause of PPG bias. Federated learning improves data diversity at the model training stage, but does not correct the hardware measurement error that corrupts the input signal before training begins. Synthetic data augmentation may improve representation statistics but cannot reproduce the physiological signal characteristics of underrepresented groups if the underlying bias is optical and physics-based. Hardware-level remediation, specifically, sensor calibration and wavelength selection validated across the Fitzpatrick scale, is necessary for any fairness intervention to be complete.

This finding aligns with Böttcher et al.'s [11] observation that data quality problems originating in device hardware cannot be fully recovered by downstream pre-processing, and with Majumder et al.'s [10] characterization of LED wavelength selection as a structural design constraint rather than a tunable parameter.

C. Deployment Considerations

A practical fairness framework for AI-enabled RPM systems using PPG data would require:

1) hardware-level validation across the full Fitzpatrick scale using arterial blood gas as ground truth rather than oximeter self-comparison
2) publicly available, demographically annotated benchmark datasets to enable reproducible cross-study evaluation
3) disaggregated subgroup performance reporting in regulatory submissions
4) explicit informed consent protocols for patients whose wearable data enters AI training pipelines
5) regulatory extension to cover sensor hardware in addition to the AI models that consume sensor output

VI. Recommendations

Based on the analysis, five targeted recommendations are proposed.

R1: Mandate Fitzpatrick-Scale Hardware Validation. The FDA should finalize binding pre-market accuracy requirements for PPG devices specifying minimum performance thresholds across all Fitzpatrick categories, using arterial blood gas as ground truth rather than oximeter self-comparison. The 2023 discussion paper framework provides a starting point; what is missing is enforcement.

R2: Establish a Public PPG Fairness Benchmark. A publicly available demographically annotated PPG dataset should be developed under NIH or equivalent sponsorship to enable reproducible, comparable evaluation of bias mitigation techniques across research groups. Without this, claims about bias reduction cannot be independently verified.

R3: Require Subgroup Performance Reporting. FDA and EU regulatory submissions for AI-enabled RPM tools should require disaggregated performance reporting by skin tone using standard fairness metrics, not aggregate accuracy alone. This applies both to the AI model and to the sensor hardware that feeds it.

R4: Extend EU AI Act Scope to Sensor Hardware. The high-risk medical AI classification under the EU AI Act should explicitly cover the sensor hardware that powers those systems, making optical calibration requirements enforceable alongside model governance obligations. Currently, a manufacturer can comply with the Act's documentation requirements while deploying sensors with known skin tone bias.

R5: Adopt Wearable Data Consent Frameworks. Institutions deploying RPM systems that feed patient PPG data into AI training pipelines should implement explicit consent protocols that disclose this use and allow opt-out without affecting clinical care, as recommended by Park [16].

VII. Conclusion

This paper traced the complete causal chain from the optical physics of wearable PPG sensors through AI training pipelines to clinical outcomes and regulatory governance. This measurement bias is not correctable by algorithmic intervention alone because it represents a signal that was never accurately captured.

The regulatory response has been inadequate. The FDA has not published pre-market mandates for this PPG measurement bias it acknowledged in 2021. The EU AI Act provides a stronger enforcement architecture, but it does not contain a PPG-specific hardware validation standard. Five critical gaps remain unaddressed: no standardized fairness benchmark, no finalized validation protocol, limited research-to-deployment evidence, weak informed consent frameworks, and unrecognized socioeconomic compounding.

Future work should prioritize the development of a public Fitzpatrick-annotated PPG benchmark dataset and prospective evaluation of fairness interventions in live clinical workflows. Next steps should also consider the design of regulatory testing protocols with defined sample sizes and performance thresholds per skin tone category. The technical and ethical path forward is clear; what remains 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] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.