AI Bias in Healthcare: Understanding the Problem and Charting a Path to Equity

Rasit Dinc

Rasit Dinc Digital Health & AI Research

Published: November 27, 2024 | AI Diagnostics

DOI: 10.5281/zenodo.17996895

Abstract

The integration of Artificial Intelligence AI into healthcare promises a revolution in diagnostics, treatment, and efficiency. From image recognition to pred...

The integration of Artificial Intelligence (AI) into healthcare promises a revolution in diagnostics, treatment, and efficiency. From image recognition to predictive risk models, AI is rapidly becoming an indispensable tool. However, this technological leap is shadowed by a critical challenge: **AI bias**. If left unaddressed, these biases threaten to deepen existing health disparities, compromise patient safety, and undermine the promise of equitable care.

The Genesis of Bias: A Systemic Problem

AI bias in healthcare is not a technical glitch but a systemic issue rooted in the data and processes used to create these systems, reflecting historical and societal inequities inadvertently encoded into the algorithms. Understanding the problem requires examining the entire AI development pipeline [1], where bias can be introduced at multiple, compounding stages:

1. Data Collection and Representation

The foundation of any AI model is its training data. If this data is unrepresentative of the real-world patient population, the resulting model will inevitably be biased. An algorithm trained predominantly on data from a single ethnic group or socioeconomic class will perform poorly when applied to underrepresented groups—a phenomenon known as **sampling bias**. A widely cited example is the performance disparity of pulse oximeters, which have been shown to overestimate blood oxygen saturation in patients with darker skin pigmentation, a bias rooted in the training data and design [2].

2. Feature Selection and Annotation

Bias can also creep in during the selection of features (variables) and the annotation process. **Implicit provider bias** can influence how data is labeled. For example, if a model is trained to predict "high-risk" patients, and the training data reflects a history where certain demographic groups were

systematically undertreated, the AI may learn to associate those demographics with lower risk, simply because the historical data did not capture the true severity of their condition [3]. A classic case involved an algorithm that used healthcare costs as a proxy for health needs; because Black patients historically incurred lower costs for the same level of illness (due to systemic barriers to access), the algorithm systematically assigned them lower risk scores, thereby reducing their access to critical care [4].

3. Model Development and Evaluation

Bias can persist even into the model's evaluation phase. If the metrics used to assess a model’s performance—such as accuracy—are not disaggregated by demographic groups (e.g., race, gender, age), a model that is highly accurate overall may still be dangerously inaccurate for a minority group. This is a failure of **developer naivety** or insufficient diligence in ensuring fairness across all subgroups [1].

The Clinical Consequences: Exacerbating Disparities

The consequences of biased AI are not abstract; they manifest as tangible harm to patients. This includes **Misdiagnosis and Delayed Treatment**, such as algorithms trained on images of light skin failing to accurately diagnose skin conditions in individuals with darker skin [5]. It also leads to **Resource Misallocation**, where biased predictive models incorrectly triage patients, perpetuating systemic health inequities. Finally, it causes an **Erosion of Trust** in both the technology and the healthcare system when patients realize the technology is failing them based on their background.

Charting a Path to Equitable AI

Addressing AI bias requires a multi-pronged approach that spans policy, technology, and ethics, demanding a commitment to **Fairness, Accountability, and Transparency (FAT)**. This involves: **Data Equity and Auditing**, where developers actively seek diverse, high-quality datasets and conduct regular, independent audits to correct representational gaps; adopting **Algorithmic Fairness Metrics** that move beyond aggregate accuracy to ensure equitable performance across all demographic subgroups; and fostering **Interdisciplinary Collaboration** by involving ethicists, social scientists, and patient advocates alongside data scientists from the initial design phase.

The challenge of AI bias is significant, but by committing to rigorous, equitable development practices, we can ensure that AI fulfills its potential as a force for good, advancing health equity rather than hindering it. For more in-depth analysis on the ethical and technical frameworks required to build trustworthy AI in medicine, the resources at www.rasitdinc.com provide expert commentary and professional insight.

**

References

[1] Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. PLOS Digital Health*, 3(11), e0000651. [https://doi.org/10.1371/journal.pdig.0000651] (https://doi.org/10.1371/journal.pdig.0000651) [2] Sjoding, M. W.,* et al. *(2020). Racial Bias in Pulse Oximetry Measurement.* The New England Journal of Medicine*, 383(25), 2478–2480. [https://doi.org/10.1056/NEJMc2029240] (https://doi.org/10.1056/NEJMc2029240) [3] Norori, N.,* et al. *(2021). Addressing bias in big data and AI for health care.* The Lancet Digital Health*, 3(12), e809-e818. [https://doi.org/10.1016/S2589-7500(21)00195-0] (https://doi.org/10.1016/S2589-7500(21)00195-0) [4] Obermeyer, Z.,* et al. *(2019). Dissecting racial bias in an algorithm used to manage the health of populations.* Science*, 366(6464), 447-453. [https://doi.org/10.1126/science.aax2342] (https://doi.org/10.1126/science.aax2342) [5] Adamson, A. S., & Norris, T. L. (2021). Race, technology, and the future of dermatology.* Journal of the American Academy of Dermatology*, 84(2), 525-527. [https://doi.org/10.1016/j.jaad.2020.08.067] (https://doi.org/10.1016/j.jaad.2020.08.067)
