The Algorithmic Divide: Ensuring Bias and Fairness in Healthcare Machine Learning Models

Rasit Dinc

Rasit Dinc Digital Health & AI Research

Published: April 23, 2025 | Medical Imaging AI

DOI: 10.5281/zenodo.17996722

Abstract

The Algorithmic Divide: Ensuring Bias and Fairness in Healthcare Machine Learning Models The integration of Machine Learning (ML) models into healthcare...

The Algorithmic Divide: Ensuring Bias and Fairness in Healthcare Machine Learning Models

The integration of **Machine Learning (ML)** models into healthcare promises a revolution in diagnostics, treatment planning, and public health management. However, this powerful technology is not without its ethical and practical challenges. Chief among these is the pervasive issue of **bias and fairness** [1]. If left unaddressed, algorithmic bias in healthcare ML can exacerbate existing health disparities, erode patient trust, and ultimately compromise the quality of care [2]. For professionals in digital health and AI, understanding and actively mitigating these biases is not merely an ethical imperative but a foundational requirement for responsible innovation.

The Sources and Types of Algorithmic Bias

Bias in ML models is rarely intentional; it is typically a reflection of the data used to train them or the way the models are designed and deployed [3]. In healthcare, the sources of bias are complex and deeply rooted in historical and systemic inequities.

1. Data-Driven Biases

The most common source of bias is the training data itself.

Representation Bias: *This occurs when the training dataset does not accurately reflect the diversity of the target population, particularly underrepresenting minority groups, women, or specific socioeconomic strata [4]. For instance, a diagnostic model trained predominantly on data from one ethnic group may perform poorly or inaccurately for others.* **Historical Bias:** This arises when the data reflects past or current societal prejudices. A model

trained on historical data where certain groups received suboptimal care or were systematically under-diagnosed for specific conditions will learn and perpetuate those same discriminatory patterns [5]. ***Measurement Bias:*** *This involves systematic errors in how data is collected or labeled. In medical imaging, for example, variations in scanner quality or image acquisition protocols across different hospitals can introduce bias that the model learns to associate with patient outcomes [6].*

2. Algorithmic and Systemic Biases

Beyond the data, bias can be introduced during the model development and deployment phases.

Algorithmic Bias: This relates to the choice of fairness metrics or the structure of the model itself. Different mathematical definitions of "fairness" (e.g., equal opportunity, equalized odds, demographic parity) can lead to different, and sometimes conflicting, outcomes for various subgroups [7]. ***Systemic Bias:*** *This is the bias that emerges when a model is deployed into a real-world clinical workflow. A model that predicts which patients would benefit from a high-cost intervention might systematically favor patients from higher-income hospitals, simply because the training data correlated high-cost intervention with better outcomes in those settings [8].*

Ethical Implications and Health Equity

The consequences of biased healthcare ML models are profound, extending beyond mere technical inaccuracy to impact fundamental issues of health equity. A landmark study demonstrated how a widely used commercial algorithm for managing the health of millions of people in the US systematically assigned lower risk scores to Black patients than to equally sick white patients, leading to fewer Black patients being referred for necessary care management programs [9].

This type of algorithmic discrimination can lead to:

Misdiagnosis and Suboptimal Treatment: Biased models can result in delayed or incorrect diagnoses for underrepresented groups, widening the gap in health outcomes. ***Erosion of Trust:*** *If patients perceive that AI systems are treating them unfairly, their trust in the healthcare system, and in the technology itself, will diminish, leading to lower adoption and compliance [2].* **Perpetuation of Inequity:** By automating and scaling historical biases, ML models risk cementing systemic discrimination into the future of healthcare delivery [5].

Strategies for Mitigation and Responsible Development

Addressing bias requires a multi-faceted approach that spans the entire ML lifecycle, from data collection to deployment and monitoring [10].

| Mitigation Strategy | Description | ML Lifecycle Stage | | :--- | :--- | :--- | | **Fairness-Aware Data Curation** | Actively audit and rebalance datasets to ensure demographic and clinical diversity. Use techniques like oversampling or synthetic data generation for underrepresented groups. | Data Pre-

processing | | **Bias Detection and Auditing** | Employ specialized fairness metrics (e.g., disparate impact, equalized odds) to systematically test model performance across different sensitive subgroups (e.g., race, gender, age). | Model Training & Evaluation | | **In-Processing Mitigation** | Integrate fairness constraints directly into the model training objective function, forcing the model to optimize for both accuracy and fairness simultaneously. | Model Training | | **Post-Processing Techniques** | Adjust the model's output or decision threshold after training to achieve a desired level of fairness for specific subgroups. | Model Deployment | | **Transparency and Explainability (XAI)** | Use Explainable AI techniques to understand *why* a model made a particular decision, making it easier to trace and correct biased behavior. | Model Evaluation & Deployment | | **Stakeholder Involvement** | Engage diverse clinical, ethical, and community stakeholders throughout the development process to define what "fairness" means in a specific clinical context. | All Stages |

The future of digital health relies on the development of ML models that are not only accurate but also equitable. By prioritizing rigorous data auditing, adopting fairness-aware modeling techniques, and ensuring continuous monitoring in real-world settings, the digital health community can build a future where AI serves as a powerful tool for reducing, rather than amplifying, health disparities [1] [10].

***

References

[1] [Fairness of artificial intelligence in healthcare: review and perspective](https://pmc.ncbi.nlm.nih.gov/articles/PMC10764412/) [2] [Ethical Implications of Algorithmic Bias in Medical AI](https://prism.sustainability-directory.com/scenario/ethical-implications-of-algorithmic-bias-in-medical-ai/) [3] [What is AI bias? Causes, effects, and mitigation strategies](https://www.sap.com/resources/what-is-ai-bias) [4] [Bias recognition and mitigation strategies in artificial intelligence in healthcare](https://pmc.ncbi.nlm.nih.gov/articles/PMC11897215/) [5] [Artificial intelligence and algorithmic bias: implications for health systems](https://pmc.ncbi.nlm.nih.gov/articles/PMC6875681/) [6] [Understanding and mitigating bias in imaging artificial intelligence](https://pubs.rsna.org/doi/abs/10.1148/rg.230067) [7] [Equity in essence: a call for operationalising fairness in machine learning for healthcare](https://pmc.ncbi.nlm.nih.gov/articles/PMC8733939/) [8] [Evaluating the impact of data biases on algorithmic fairness in clinical machine learning models](https://academic.oup.com/jamiaopen/article/8/5/ooaf115/8269322) [9] [Dissecting racial bias in an algorithm used to manage the health of millions of people](https://www.science.org/doi/10.1126/science.aax2342) [10] [Bias in AI-based models for medical applications: challenges and mitigation strategies](https://www.nature.com/articles/s41746-023-00858-z)