When Bias Becomes Knowledge: How Sociodemographic Inequities Shape Medical AI

Om M. Patel[1]

[1] Faculty of Health Sciences, McMaster University, Hamilton, ON, Canada

Correspondence: patelo31@mcmaster.ca

Date Published: April 30, 2026
DOI: https://doi.org/10.18192/UOJM.V16iS1.7834

Artificial intelligence (AI), especially with the recent advancements of large-language models (LLMs), has become an integral part of clinical decision-making, affecting triage, risk stratification, diagnosis, and treatment selection.[1] These tools are typically presented as ways to improve efficiency and objectivity within healthcare. However, emerging research suggests that healthcare-related artificial intelligence does something more consequential and troubling: it takes pre-existing biases in clinical practice and translates them into a seemingly "objective" clinical judgment.[2]

Recently, a 2025 study evaluated nine different LLMs, and demonstrated that even when the same clinical data is inputted, triage priority, treatment recommendations, and even the mental health assessments differed solely on the basis of race or gender-based demographic indicators.[3] As opposed to correcting inequities, large-language models seem to legitimize them, under the guise of objectivity.[3]

The majority of the evaluation metrics pertaining to medical AI systems focuses primarily on technical performance, measuring values such as sensitivity, specificity, area under the receiver operating characteristic curve (AUC-ROC), and F1-scores (a composite measure of recall and precision).[4] These metrics address the wrong question entirely. The core issue here is not about the performance of these AI systems, but about the type of knowledge that is being learned and reproduced by them.[5]

The root cause for AI biases is well understood. Decades of research document systematic disparities in care. For instance, in comparable injuries, Black patients are significantly less likely than White patients to receive opioid analgesia in emergency department settings.[6] Women with acute coronary syndrome experience longer times to diagnosis and are less likely to receive guideline-consistent treatment.[7] Black patients are more likely to be diagnosed with schizophrenia and less likely to be diagnosed with mood disorders compared to White patients presenting with similar symptoms.[8] These inequities remain significant after adjusting for clinical severity and comorbidities, suggesting that they cannot be explained by medical factors alone. When such clinical decisions are treated as ground truth by machine-learning systems, historical inequities are learned, reinforced, and then propagated throughout the healthcare system.[5]

Bias also enters healthcare through clinical documentation. Studies of electronic health records (EHR) show that symptoms of Black patients are more likely to be documented in ways that suggest non-compliance or exaggeration, and symptoms of women are more likely to be attributed to anxiety or stress.[9] Medical documentation is the primary training input for many AI models, meaning what is considered neutral training data is already plagued with social bias.[5]

Machine learning algorithms trained on such data cannot differentiate between medically relevant information and systemic inequity. Their task is simply to identify statistical patterns which predict future outcomes. When historical practice has been unequal, the algorithms encode these disparities as a normative pattern. From the model's perspective, unequal care is not a problem to be solved, but the very process from which it learns. Bias has thus, become knowledge.[10]

The true danger of algorithmic bias lies in its authority. Unlike clinicians who can reflect and change behavior, algorithms codify historical patterns as fixed parameters, reproducing them consistently at scale without capacity

for self-correction. Their recommendations are often perceived as data-driven, but this perception is misleading.[5] This contributes to automation bias, where clinicians defer to algorithmic recommendations even when they conflict with clinical judgment.[11] Bias that once emerged episodically in bedside encounters now operates continuously at population scale, shaping care across institutions in ways that are systematic, persistent, and far more difficult to rectify.

No single solution can fully address this. From a technical perspective, evaluation criteria need to shift from purely performance-focused validation to evaluation methods that interrogate bias directly.[12] A promising method would be counterfactual stress testing, where models are tested on the same set of clinical cases with varying social cues such as names or pronouns. This method directly assesses whether models are biased towards demographic identifiers rather than medical data while making recommendations.[13] Adding such tests to the pre-deployment validation process would enable institutions to identify bias mechanisms that are normally missed by standard validation metrics.[14]

However, technical solutions alone are not sufficient. As long as AI models are trained on clinical documents that are reflective of inequitable care, models will inevitably learn inequity as the ground truth.[15] This highlights the need to intervene at the data curation and annotation phase. Instruction tuning datasets with balanced representation of clinical presentations across sociodemographic groups can minimize the strength of associations learned by models between certain conditions/behaviors and sociodemographic groups.[10] Furthermore, data augmentation techniques, which involve the use of counterfactuals such as swapping demographic information and keeping medical content constant, can also be employed to break associations learned by models between identity and healthcare needs. While these techniques cannot completely remove inequity from model representations, they can minimize its salience.[16]

Individual stakeholders also play a key role. Medical students should begin to see AI literacy as a core clinical competency, seeking formal education on algorithmic bias. Clinicians should continue to critically appraise AI-generated recommendations, especially when they conflict with clinical judgement. Researchers and developers should also employ equity-centered validation into model design from the start, rather than treating bias audits as an afterthought.

Although AI has tremendous potential to revolutionize healthcare, one of its greatest dangers is in hardwiring such invisible biases as “knowledge” that informs care. Several challenges lie ahead. To tackle this, we need to mitigate bias in training data, rigorously test performance across subgroups, and build regulatory frameworks that keep equity enforced beyond initial model deployment. The future is promising however. Medical regulators are starting to acknowledge this need. For example, the U.S. Food and Drug Administration (FDA) new 2025 draft guidance on AI-enabled medical devices emphasizes bias mitigation.[17] The medical-AI landscape is advancing fast, and awareness among clinicians, researchers, and policymakers is growing. Done right, AI can do more than streamline care; it can push medicine toward something more equitable and transparent. That future is possible and worth pursuing.

REFERENCES

1. Maity S, Saikia MJ. Large Language Models in Healthcare and Medical Applications: A Review. Bioengineering (Basel) 2025; 12: 631.
2. Straw I. The automation of bias in medical Artificial Intelligence (Ai): Decoding the past to create a better future. Artificial Intelligence in Medicine 2020; 110: 101965.
3. Omar M, Soffer S, Agbareia R, et al. Sociodemographic biases in medical decision making by large language models. Nat Med 2025; 31: 1873–1881.
4. Hicks SA, Strümke I, Thambawita V, et al. On evaluation metrics for medical applications of artificial intelligence. Sci Rep 2022; 12: 5979.
5. Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. PLOS Digit Health 2024; 3: e0000651.
6. Jarman AF, Hwang AC, Schleimer JP, et al. Racial Disparities in Opioid Analgesia Administration Among Adult Emergency Department Patients with Abdominal Pain. West J Emerg Med 2022; 23: 826–831.
7. Lunova T, Komorovsky R, Klishch I. Gender Differences in Treatment Delays, Management and Mortality among Patients with Acute Coronary Syndrome: A Systematic Review and Meta-analysis. Curr Cardiol Rev 2023; 19: e300622206530.
8. Merritt CC, Halverson TF, Elliott T, et al. Racial Disparities and Predictors of Functioning in Schizophrenia. Am J Orthopsychiatry 2023; 93: 177–187.
9. Ivy ZK, Hwee S, Kimball BC, et al. Disparities in Documentation: Evidence of Race-Based Biases in the Electronic Medical Record. J Racial Ethn Health Disparities 2025; 12: 3294–3300.
10. Rajkomar A, Hardt M, Howell MD, et al. Ensuring Fairness in Machine Learning to Advance Health Equity. Ann Intern Med 2018; 169: 866–872.
11. Kücking F, Hübner U, Przysucha M, et al. Automation Bias in AI-Decision Support: Results from an Empirical Study. Stud Health Technol Inform 2024; 317: 298–304.
12. Char DS, Shah NH, Magnus D. Implementing Machine

Learning in Health Care — Addressing Ethical Challenges. N Engl J Med 2018; 378: 981–983.
13. Chakradeo K, Huynh I, Balaganeshan SB, et al. Navigating fairness aspects of clinical prediction models. BMC Med 2025; 23: 567.
14. Chen RJ, Wang JJ, Williamson DFK, et al. Algorithm fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng 2023; 7: 719–742.
15. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019; 366: 447–453.
16. Mehrabi N, Morstatter F, Saxena N, et al. A Survey on Bias and Fairness in Machine Learning. ACM Comput Surv 2022; 54: 1–35.
17. FDA Issues Comprehensive Draft Guidance for Developers of Artificial Intelligence-Enabled Medical Devices. FDA, https://www.fda.gov/news-events/press-announcements/fda-issues-comprehensive-draft-g uidance-developers-artificial-intelligence-enabled-medical-devices (2025, accessed 30 January 2026).

Conflicts of Interest Disclosure
There are no conflicts of interest to declare.