Quantifying Structural Healthcare Disparities through Fairness-Aware Large Language Models Integrating Multi-Modal Electronic Health Records and Socioeconomic Determinants

Julian Thorne
Department of Sociology and Public Health, Temple University
jthorne@temple.edu

Timothy Ellison
Department of Systems and Industrial Engineering, University of Arizona
t.ellison@arizona.edu

Colin Callahan
Department of Electrical and Computer Engineering, Iowa State University
colin.callahan@iastate.edu

Abstract

Structural healthcare disparities remain a persistent challenge within global health systems, often exacerbated by the historical biases embedded in clinical data and medical decision-making processes. As Large Language Models (LLMs) increasingly permeate clinical workflows, there is an urgent need to ensure these systems do not merely replicate existing inequities but actively work to quantify and mitigate them. This paper proposes a systemic framework for integrating fairness-aware LLMs with multi-modal Electronic Health Records (EHRs) and Social Determinants of Health (SDoH) to provide a granular quantification of structural disparities. By synthesizing unstructured clinical notes, longitudinal diagnostic data, and socioeconomic indicators—such as housing stability, transportation access, and neighborhood-level deprivation indices—the proposed architecture identifies latent patterns of systemic neglect and diagnostic bias. We provide a deep analytical discussion on the structural trade-offs between model interpretability, predictive accuracy, and algorithmic fairness. Furthermore, the research explores the socio-technical dimensions of deployment, emphasizing the role of algorithmic governance, data sovereignty for marginalized communities, and the policy implications of using AI as a tool for institutional audit. Our findings suggest that while LLMs possess the potential to uncover deep-seated disparities, their implementation must be grounded in a robust infrastructure of fairness-aware constraints and cross-disciplinary oversight. This paper provides a comprehensive blueprint for leveraging advanced artificial intelligence to foster a more equitable, transparent, and resilient healthcare infrastructure.

Keywords
Healthcare Disparities, Fairness-Aware AI, Large Language Models, Electronic Health Records, Social Determinants of Health, Algorithmic Governance, Socio-Technical Infrastructure.

1. Introduction

The promise of artificial intelligence in medicine has historically focused on the optimization of individual patient outcomes through enhanced diagnostic precision and personalized treatment regimens. However, this individual-centric focus often ignores the broader structural contexts in which health and disease are produced. Healthcare systems do not operate in a vacuum; they are embedded within socio-technical infrastructures that carry the weight of historical exclusions, economic stratification, and geographic inequities. Structural disparities—defined as the systematic differences in health access, quality, and outcomes across demographic and socioeconomic groups—are frequently encoded into the very data used to train clinical algorithms. Without intentional intervention, the deployment of Large Language Models (LLMs) in clinical settings risks automating and scaling these disparities under the guise of objective computation.

This paper proposes a paradigm shift toward "structural quantification" through the lens of fairness-aware LLMs. Unlike traditional predictive models that rely on tabular data, LLMs are uniquely capable of processing the vast, unstructured narratives found in clinical notes and socioeconomic assessments. These narratives often contain the "hidden" signals of disparity—subtle linguistic biases in how providers describe certain patient populations, or the indirect mention of environmental stressors that are omitted from formal diagnostic codes. By integrating these multi-modal Electronic Health Records (EHRs) with explicit socioeconomic determinants, we can develop a more holistic understanding of how systemic factors influence clinical trajectories.

The core of our investigation lies in the systemic architecture required to support such an audit. We explore the trade-offs involved in balancing the predictive power of multi-modal models with the ethical requirement of "algorithmic fairness." This involves not only the technical implementation of debiasing techniques but also a rigorous analysis of the governance structures necessary to oversee these models in practice. As we move toward a future where AI acts as a fundamental layer of the healthcare infrastructure, it is imperative that we view these models not just as clinical assistants, but as critical instruments for institutional self-reflection and policy reform.

2. Conceptual Foundations of Structural Disparity and AI Bias

To quantify structural healthcare disparities, one must first recognize that the Electronic Health Record is not a neutral mirror of biological reality; it is a socio-technical artifact. It reflects the interactions between patients and providers, the limitations of insurance coverage, and the cultural competence of the medical institution. When LLMs are trained on this data, they are essentially learning a "history of medical practice," including its flaws. Structural

bias enters the system at multiple points: from the initial point of access—where marginalized groups may be systematically underrepresented in data—to the point of documentation, where different linguistic styles may lead to disparate diagnostic outcomes.

Fairness-aware LLMs represent a specialized branch of artificial intelligence designed to identify and correct these latent biases. The conceptual challenge lies in defining "fairness" in a way that is clinically relevant and mathematically robust. Traditional metrics like demographic parity or equalized odds are often insufficient in a medical context where biological differences and legitimate clinical needs must be distinguished from systemic bias. We propose a framework of "counterfactual fairness," where the system evaluates whether a clinical recommendation would have remained consistent if a sensitive attribute—such as race or neighborhood—were changed while holding all other clinical variables constant. This approach allows the model to act as a diagnostic probe, identifying specific nodes in the clinical workflow where structural factors override biological indicators.

The integration of Social Determinants of Health (SDoH) is essential for grounding these models in reality. Socioeconomic determinants—including income volatility, educational attainment, and environmental exposures—account for a significant portion of health variance but are traditionally treated as "noise" in clinical modeling. By treating SDoH as a primary modality alongside clinical data, fairness-aware LLMs can provide a causal analysis of disparity. This allows the system to differentiate between a patient who is "non-compliant" due to behavioral choices and a patient whose clinical trajectory is hindered by structural barriers like food insecurity or lack of reliable transportation. This conceptual shift moves the burden of "health improvement" from the individual back to the systems responsible for their care.

3. Architecture for Multi-Modal Data Integration and Fairness

The systemic architecture for quantifying disparities must be multi-modal by design, capable of ingesting and aligning diverse data streams into a unified representation. We propose a "tri-modal" architecture that integrates longitudinal EHR data (laboratory results, vital signs), unstructured clinical text (physician notes, discharge summaries), and geographic-socioeconomic indices (Area Deprivation Index, housing data). The technical challenge involves "semantic alignment," ensuring that the model understands the relationship between a patient’s rising hemoglobin A1c levels and their residence in a known "food desert." This requires the use of cross-modal transformer architectures that can attend to both clinical and environmental signals simultaneously.

Within this architecture, the "Fairness-Aware Core" operates as a secondary optimization layer. Instead of purely optimizing for the minimization of loss on a diagnostic task, the model is trained with a multi-objective loss function that includes a "disparity penalty." This penalty is triggered whenever the model’s internal representations show excessive correlation between sensitive socioeconomic attributes and predicted clinical outcomes that cannot be explained by biological variables. This architectural constraint forces the LLM to seek features that are more robust and less prone to echoing historical bias. Furthermore, we

advocate for the use of "explainable modules" that allow clinicians to see which socioeconomic or clinical factors were most influential in a specific quantification of disparity.

Infrastructure deployment for such systems must be sensitive to the realities of data siloing in healthcare. We propose a "federated fairness" model, where the LLM is trained across multiple institutions without moving sensitive patient data. Each hospital or clinic trains a localized version of the model that reflects its specific demographic mix, and only the "fairness gradients"—the updates needed to improve the model's objectivity—are shared across the network. This decentralized approach protects patient privacy and data sovereignty while allowing for a large-scale, national-level audit of healthcare inequities. This infrastructure ensures that the system is not only powerful but also resilient to the variations in data quality and documentation styles found across different healthcare regions.

4. Structural Trade-offs: Interpretability vs. Predictive Precision

The implementation of fairness-aware LLMs involves a constant negotiation between competing technical and ethical objectives. The most prominent trade-off is between "predictive precision" and "model interpretability." In many medical applications, the highest-performing models are "black boxes" that utilize complex, non-linear relationships to achieve accuracy. However, in the context of structural disparity, a model that cannot explain why it has identified a particular bias is of limited use for policy reform. To address this, we must accept a marginal "precision tax"—a slight reduction in raw predictive power—in exchange for architectures that allow for causal tracing of disparate outcomes.

Another critical trade-off exists between "group fairness" and "individual fairness." Group fairness aims to ensure that outcomes are balanced across demographic aggregates (e.g., ensuring equal diagnostic rates for all racial groups). Individual fairness, however, demands that similar individuals receive similar treatments regardless of their group membership. In a healthcare system shaped by deep structural inequities, these two objectives can often conflict. For instance, a model designed to achieve group parity might overlook unique individual needs within a sub-population. Our systemic framework utilizes a "hierarchical fairness" approach, where group-level metrics are used to identify large-scale institutional failures, while individual-level counterfactuals are used to audit specific clinical decisions.

Sustainability is a further structural consideration. The computational cost of training and running multi-modal LLMs with fairness constraints is significant. In many resource-constrained clinical settings, the infrastructure required to host such models may be unavailable. We argue for the development of "distilled" clinical models—smaller, highly efficient versions of LLMs that are pre-conditioned with fairness constraints but require less compute to run at the point of care. This ensures that the ability to quantify and mitigate disparities is not restricted to elite academic medical centers but is an accessible tool for rural and community health systems that often serve the most vulnerable populations.

5. Deployment Challenges and Socio-Technical Robustness

Deploying fairness-aware LLMs in clinical environments requires navigating a complex landscape of institutional inertia and technical fragmentation. One major challenge is the "documentation-disparity feedback loop." If a healthcare system systematically under-documents the needs of certain groups, the LLM will inherit this "silence" as a lack of need. A robust system must therefore include "uncertainty quantification," where the model explicitly flags instances where it lacks sufficient socioeconomic or clinical context to make a fair determination. This forces a human-in-the-loop intervention, prompting the provider to collect more detailed social histories.

Robustness in this context also refers to the system's ability to handle "concept drift"—the changing nature of socioeconomic stressors and clinical practices over time. For example, the drivers of healthcare disparity during a global pandemic are significantly different from those during a period of economic stability. A fairness-aware architecture must include "continuous audit" protocols, where the model's performance on marginalized sub-populations is monitored in real-time. If the model's fairness metrics begin to degrade—indicating that new biases are entering the data—the system must trigger a re-training or calibration cycle. This ensures that the AI remains a reliable tool for equity even as the external environment evolves.

Furthermore, the integration of these models into the clinical workflow must be handled with extreme care to avoid "automation bias," where providers uncritically follow the AI's suggestions. Socio-technical robustness requires that the LLM's output is presented not as a final judgment, but as a "disparity alert" or a "contextual summary." For instance, when a physician is reviewing a patient with frequent emergency room visits, the LLM might provide a summary stating: "Clinical records suggest recurrent asthma exacerbations; however, environmental data indicates housing instability in a high-pollution zip code, suggesting that clinical intervention alone may be insufficient." This shifts the provider's focus from a purely biological symptom to a structural cause, effectively using AI to expand the clinician's diagnostic horizon.

6. Algorithmic Governance and the Ethics of Institutional Audit

The use of AI to quantify structural disparities moves the technology into the realm of institutional audit and legal liability. If an LLM identifies that a specific hospital department systematically provides lower-quality care to a protected group, the institution faces significant legal and reputational risks. Consequently, algorithmic governance is not merely a technical requirement but a fundamental part of the system's social license to operate. We propose a governance model based on "independent algorithmic oversight," where the fairness audits of clinical LLMs are performed by third-party boards consisting of ethicists, community advocates, and data scientists.

Governance also extends to the "ownership of the narrative." Marginalized communities have historically been the subjects of medical research without having a say in how their data is interpreted. A fairness-aware infrastructure must prioritize "community data sovereignty." This involves involving community stakeholders in the definition of the fairness metrics

themselves. For example, a community in an urban area may prioritize "environmental justice" metrics, while a rural community may focus on "transportation equity." By allowing communities to help shape the model's objective functions, we ensure that the "quantification of disparity" reflects the lived experience of the patients rather than just the assumptions of the developers.

Furthermore, we must address the "ethics of intervention." Simply identifying a disparity is insufficient; the system must be linked to actionable policy levers. If an LLM quantifies a structural barrier, there must be a clear pathway to address it—whether through the referral to social services, the adjustment of hospital staffing, or the lobbying for neighborhood-level environmental changes. Algorithmic governance must therefore bridge the gap between the "data science" of the AI and the "social work" of the healthcare system. Without this bridge, the AI risks becoming a "passive observer" of suffering, documenting inequity without providing the tools for its dismantling.

7. Policy Implications and the Future of Equitable Healthcare

The integration of fairness-aware LLMs into the national healthcare infrastructure has profound implications for health policy and insurance regulation. At the federal level, regulators could utilize these models to establish "equity benchmarks" for healthcare providers. Hospitals that demonstrate significant, AI-quantified reductions in structural disparities could be incentivized through value-based payment models. Conversely, persistent, unaddressed disparities identified by fairness-aware audits could trigger regulatory interventions. This would transform "health equity" from a vague aspirational goal into a measurable, reportable metric of institutional performance.

On the level of insurance and reimbursement, policy must evolve to recognize "structural intervention" as a reimbursable clinical activity. If an LLM identifies that a patient's health trajectory is being derailed by a socioeconomic determinant, the healthcare system should be incentivized to treat that determinant with the same urgency as a biological pathogen. This requires a shift from a "fee-for-service" model to a "population health" model, where the success of the system is measured by its ability to flatten the disparity curve. Policy-makers must also ensure that the data used for these models—particularly sensitive socioeconomic data—is protected by stringent privacy laws to prevent its misuse by entities seeking to discriminate against high-risk populations.

Looking toward the future, we envision a "Global Health Equity Network," where fairness-aware LLMs provide real-time monitoring of disparities across diverse international health systems. This would allow for the cross-national comparison of structural barriers and the identification of "best-practice" infrastructures that have successfully mitigated historical inequities. The ultimate goal is an AI-augmented healthcare system that is "color-blind" to prejudice but "color-conscious" to the structural realities of the human condition. By quantifying the invisible, we can begin the hard work of making the healthcare system truly universal.

8. Forward-Looking Perspectives: Toward a Causal Healthcare AI

The next frontier in healthcare AI lies in the transition from "correlation-based" models to "causal" models. While current LLMs are excellent at identifying patterns of disparity, they often struggle to differentiate between coincidental correlations and true causal drivers. Future architectures will likely integrate "causal graphs"—pre-defined maps of how socioeconomic and biological factors interact—to provide more robust quantifications of disparity. This would allow an LLM to state with confidence: "The disparity in cardiovascular outcomes for this group is 60% attributable to lack of pharmacy access and 40% to historical diagnostic delay." Such granular causal analysis would be a revolutionary tool for public health planning.

We also anticipate the rise of "generative fairness," where LLMs are used to create "fair synthetic data." In many cases, the lack of data on marginalized groups prevents the training of accurate models. Generative AI could be used to create high-fidelity, privacy-preserving synthetic patient cohorts that "fill in the gaps" of the medical record, allowing for more robust testing of fairness constraints. However, this must be balanced against the risk of hallucinating "perfectly fair" worlds that ignore the persistent realities of structural oppression. The tension between "modeling the world as it is" and "modeling the world as it should be" will be a central theme of the next decade of AI research.

Finally, the role of the patient in this infrastructure will become more active. We foresee the development of "Personal Health Advocates"—patient-facing LLMs that help individuals navigate the structural barriers identified by the clinical models. A patient's AI assistant might alert them to a potential bias in a provider's recommendation or suggest community resources based on the structural stressors it identifies. This democratization of AI ensures that the "quantification of disparity" is not just a top-down institutional process but a bottom-up tool for patient empowerment and advocacy. The systemic integration of fairness, multi-modality, and socioeconomic context is the essential first step toward this more equitable future.

9. Conclusion

Structural healthcare disparities are not inevitable; they are the result of specific, traceable systemic failures. This paper has argued that fairness-aware Large Language Models, when integrated with multi-modal EHRs and socioeconomic determinants, provide a powerful new instrument for the quantification and mitigation of these failures. By treating the clinical narrative and the social environment as primary clinical data, we can uncover the latent biases that have historically been hidden in the "white space" of the medical record.

However, the path to equitable AI is not purely technical. It requires a robust socio-technical infrastructure grounded in algorithmic governance, community sovereignty, and bold policy reform. The trade-offs between precision and fairness must be navigated with an explicit commitment to social justice, ensuring that AI is used to dismantle, rather than reinforce, the barriers to health. As AI becomes an invisible layer of the healthcare system, it must be tasked with the most difficult work of all: auditing the institutions that created it. By holding a mirror to our systemic flaws, fairness-aware LLMs offer a path toward a healthcare infrastructure that is truly resilient, transparent, and just.

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308-318.

2. Adler, N. E., & Stewart, J. (2010). Health disparities across the lifespan: Meaning, methods, and mechanisms. Annals of the New York Academy of Sciences, 1186(1), 5-23.

3. Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. California Law Review, 104, 671.

4. Benjamin, R. (2019). Race After Technology: Abolitionist Tools for the New Jim Code. Polity.

5. Bommasani, R., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

6. Braveman, P., & Gottlieb, L. (2014). The social determinants of health: It's time to consider the causes of the causes. Public Health Reports, 129(1_suppl2), 19-31.

7. Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical strategy? Journal of the American Medical Association (JAMA), 321(16), 1549-1550.

8. Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. arXiv preprint arXiv:1808.00023.

9. Dastin, J. (2018). Amazon scraps AI recruiting tool that showed bias against women. Reuters.

10. Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, 214-226.

11. Eubanks, V. (2018). Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Press.

12. Gebru, T., et al. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86-92.

13. Ghassemi, M., Naumann, T., Schulam, P., Beam, A. L., Chen, I. Y., & Ranganath, R. (2020). A review of challenges and opportunities in machine learning for health. AMIA Joint Summits on Translational Science Proceedings, 2020, 191.

14. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29.

15. Hoffman, K. M., Trawalter, S., Axt, J. R., & Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. Proceedings of the National Academy of Sciences, 113(16), 4296-4301.

16. Johnson, A. E., et al. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3(1), 1-9.

17. Kusner, M. J., Loftus, J., Russell, C., & Silva, R. (2017). Counterfactual fairness. Advances in Neural Information Processing Systems, 30.

18. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1-35.

19. Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press.

20. Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

21. Pasquale, F. (2015). The Black Box Society: The Secret Algorithms That Control Money and Information. Harvard University Press.

22. Rajkomar, A., et al. (2018). Scalable and accurate deep learning with electronic health records. npj Digital Medicine, 1(1), 1-10.

23. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine, 169(12), 866-872.

24. Sun, M., et al. (2022). Health equity and machine learning in medicine. JAMA Network Open, 5(3), e224403.

25. Vaswani, A., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

26. Vyas, A. N., Eisenstein, L. G., & Jones, D. S. (2020). Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. New England Journal of Medicine, 383(9), 874-882.

27. Wiens, J., et al. (2019). Do no harm: A roadmap for responsible machine learning for health. Nature Medicine, 25(9), 1337-1340.

28. Williams, D. R., & Mohammed, S. A. (2013). Racism and health I: Pathways and scientific evidence. American Behavioral Scientist, 57(8), 1152-1173.

29. Yue, Y., Khanal, A., Lyu, T., Weissman, S., & Liang, C. (2025, May). EHR Phenotyping Methods for Measuring Treatment Adherence Among People Living With HIV in All of Us: Towards Disparities and Inequalities in HIV Care Continuum. In AMIA Annual Symposium Proceedings (Vol. 2024, p. 1294).

30. Zuboff, S. (2019). The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs.