A Distributed Cloud-Edge Infrastructure for Equitable Healthcare: Scaling Privacy-Preserving LLMs to Identify Regional Care Disparities

Elena Rodriguez
Department of Public Health Sciences, University of North Carolina at Charlotte
erodriguez@uncc.edu

Abstract

The persistence of regional healthcare disparities remains a critical challenge for global public health, exacerbated by the fragmentation of medical data and the limitations of centralized analytical models. While Large Language Models (LLMs) offer transformative potential for synthesizing unstructured clinical notes and identifying social determinants of health, their deployment is often hindered by stringent privacy regulations and the computational bottleneck of centralizing sensitive patient records. This paper proposes a distributed cloud-edge infrastructure designed to facilitate equitable healthcare by scaling privacy-preserving LLMs across geographically dispersed clinical environments. We introduce a tiered architectural framework that leverages edge computing to perform local, privacy-compliant data processing, while utilizing a secure cloud orchestrator for global disparity synthesis. Our analysis focuses on the system-level trade-offs between local inference latency, global model coherence, and the robust enforcement of patient confidentiality. We examine the socio-technical dimensions of this infrastructure, including algorithmic fairness in underrepresented regions, the environmental sustainability of distributed medical AI, and the policy implications for multi-jurisdictional healthcare governance. By integrating federated learning protocols with hardware-verified security, the proposed framework provides a scalable roadmap for identifying and mitigating care inequities without compromising data sovereignty. The discussion concludes with a forward-looking perspective on the ethics of automated health equity assessments and the evolving regulatory landscape surrounding decentralized medical intelligence.

Keywords

Distributed Systems, Health Equity, Edge Computing, Large Language Models, Privacy-Preserving AI, Socio-Technical Infrastructure, Healthcare Policy.

1. Introduction

The pursuit of health equity requires a granular understanding of how geography, infrastructure, and socioeconomic status intersect to produce divergent clinical outcomes. Despite advancements in medical technology, regional care disparities persist, often hidden within the vast volumes of unstructured clinical data that traditional statistical models fail to penetrate. Large Language Models (LLMs) have emerged as a powerful tool for decoding these complexities, capable of extracting nuanced insights from physician notes, patient

histories, and social intake forms. However, the application of LLMs to regional disparity identification faces a fundamental architectural paradox: the need for massive, diverse datasets to train accurate models conflicts with the legal and ethical imperative to keep patient data localized and secure.

Centralized healthcare AI infrastructures, while computationally efficient, create significant risks of data breaches and often fail to account for the unique linguistic and clinical contexts of regional healthcare providers. Furthermore, the reliance on centralized repositories often excludes smaller, rural, or under-resourced clinics that lack the high-bandwidth connectivity or technical infrastructure to export large datasets. This exclusion leads to "data deserts," where the very populations most affected by healthcare disparities are omitted from the models designed to identify them. This paper proposes a distributed cloud-edge infrastructure as a socio-technical intervention to bridge this gap, ensuring that health equity analysis is both inclusive and privacy-preserving.

The proposed system-level architecture distributes the computational load of LLM inference and disparity identification across a continuum of edge nodes—located within regional clinics—and a centralized cloud backbone. This decentralized approach allows sensitive patient data to remain behind local firewalls, while the cloud layer synthesizes non-sensitive, high-level insights across regions. By focusing on the structural trade-offs of this deployment, we explore how distributed systems can be leveraged to uphold fairness and robustness in healthcare. The integration of such technology into the healthcare ecosystem is not merely a technical challenge but a policy-driven necessity that demands a rigorous examination of governance, sustainability, and the ethical implications of autonomous disparity detection.

2. Conceptualizing Distributed Clinical Intelligence

Clinical intelligence has traditionally been viewed as a centralized resource, where data is pulled into a common repository for retrospective analysis. However, the rise of edge computing necessitates a shift toward a "distributed intelligence" paradigm. In this model, the edge is not merely a data collector but a site of active cognition. For healthcare equity, this means that the identification of care gaps occurs at the point of care, allowing for immediate contextualization within the local environment. A distributed infrastructure recognizes that a clinical note written in an Appalachian rural clinic carries different semantic weight than one written in a metropolitan teaching hospital, even if the clinical symptoms are identical.

The synergy between edge-based clinical processing and cloud-based synthesis allows for a multi-scale understanding of healthcare. At the local level, edge nodes run distilled LLMs that identify immediate disparities—such as a lack of follow-up for chronic conditions within a specific zip code. These nodes then communicate abstract representations of these findings to a cloud orchestrator. The cloud layer does not see the patient's name or specific medical history; instead, it sees emerging patterns of care gaps across multiple regions. This dual-layered reasoning provides a more robust defense against "model drift," as the global model is constantly informed by diverse, localized clinical realities without the risk of data leakage.

From a systems perspective, this distributed clinical intelligence is inherently socio-technical. It requires a alignment between the technical protocols of federated learning and the social protocols of clinical trust. For a regional clinic to participate in an equity-identification network, the system must provide verifiable guarantees of data sovereignty. The infrastructure must also be resilient to the "cold start" problem, where under-resourced clinics may have sparse data. By utilizing transfer learning from the global cloud model, edge nodes can maintain high diagnostic and analytical accuracy even in regions with smaller patient populations, ensuring that health equity assessments are not biased toward high-volume urban centers.

3. Tiered Cloud-Edge Architecture and Global Orchestration

The physical and logical architecture of the proposed system is built on a tiered structure designed to maximize both privacy and analytical throughput. The Tier 1 Edge layer resides within the clinical firewall. These are high-performance localized compute nodes capable of hosting quantized LLMs. Their primary function is the ingestion and de-identification of unstructured clinical streams. By performing the heaviest NLP tasks—such as entity recognition and social determinant extraction—locally, the edge layer minimizes the bandwidth requirements and eliminates the need to transmit raw Protected Health Information (PHI) over public networks. This architectural choice is driven by the legal constraints of HIPAA and GDPR, transforming compliance into a structural feature of the system.

Tier 2 represents the Regional Aggregator layer, which serves as a secure bridge between clusters of edge nodes and the global cloud. In many regional healthcare networks, data sharing is permitted within a specific hospital system but restricted outside of it. The Regional Aggregator facilitates the first level of synthesis, identifies intra-network disparities, and manages the distribution of model updates. This layer acts as a cache for model weights, ensuring that even if connectivity to the global cloud is intermittent—a common occurrence in rural regions—the edge nodes can continue to function using the most recent localized intelligence. This hierarchical caching is a critical structural trade-off that prioritizes system availability over perfect global synchronization.

The Tier 3 Global Cloud layer serves as the "equity orchestrator." This is where the cross-modal synthesis of regional data occurs. The cloud engine utilizes high-level semantic embeddings sent from the aggregators to build a comprehensive map of national or global healthcare disparities. The cloud layer is also responsible for the "fairness audit" of the global LLM. By comparing the insights generated from different regions, the orchestrator can identify if the model is systematically underperforming in specific cultural or linguistic contexts. This tiered architecture ensures that the system is scalable, moving from local clinical insights to regional aggregators and finally to a global perspective on health equity, all while maintaining a rigid privacy-preserving boundary at the edge.

4. Privacy-Preserving LLM Scaling and Federated Synthesis

Scaling LLMs in a healthcare environment requires overcoming the immense computational

cost of inference while maintaining strict isolation between datasets. Our framework utilizes a federated synthesis protocol where model training and disparity identification are decoupled. In this paradigm, the global LLM is not "trained" on patient data in the traditional sense. Instead, the edge nodes perform "local reasoning" on PHI and then transmit only the logical conclusions and non-sensitive gradients to the cloud. This ensures that the global model learns the logic of care disparities without ever being exposed to the underlying patient records.

To further enhance privacy, we integrate differential privacy techniques into the gradient exchange. By adding controlled statistical noise to the data transmitted to the cloud, the system ensures that it is mathematically impossible to reconstruct an individual patient's record from the global model's updates. This introduces a structural trade-off: higher levels of privacy noise can slightly degrade the precision of the model. However, in the context of identifying regional disparities, where the goal is to identify population-level trends rather than individual diagnoses, this marginal loss in precision is an acceptable compromise for the gain in system-wide trust and regulatory compliance.

Furthermore, the framework employs secure multi-party computation (SMPC) for the most sensitive comparison tasks. If two regional clinics wish to determine if they are seeing similar disparities without revealing their specific patient demographics to each other, SMPC allows them to compute a joint analytical result without either party ever seeing the other's input. This level of privacy-preserving scaling is essential for fostering collaboration between competing healthcare providers, turning the infrastructure into a "neutral ground" for health equity research. By building privacy into the communication protocol rather than treating it as a post-processing step, the system achieves a level of robustness that centralized models cannot match.

5. Infrastructure Robustness and Adversarial Resilience

In the context of healthcare, system robustness is synonymous with patient safety and analytical integrity. A distributed infrastructure for health equity must be resilient to both accidental node failure and intentional adversarial attacks. Edge nodes in remote areas are often subject to hardware malfunctions, power instability, or cyber-vulnerabilities. Our architecture addresses this through a "decentralized health-check" mechanism. Each node in the network is monitored by its peers; if a node begins to produce anomalous disparity reports or ceases to respond, the regional aggregator automatically re-routes its workload and alerts local administrators. This self-healing property is vital for maintaining the continuity of care equity assessments in unstable environments.

Adversarial resilience is particularly critical when using LLMs to drive policy decisions. There is a risk of "narrative poisoning," where a malicious actor—perhaps attempting to hide poor performance in a specific region—could inject biased clinical notes into the edge ingestion stream to mislead the equity model. To mitigate this, we employ a multi-agent validation layer. Before a local insight is accepted into the global synthesis, it must be cross-verified by independent "auditor agents" that check for semantic consistency against historical trends and neighboring regional data. This decentralized verification process

ensures that the global equity map is not skewed by isolated instances of data corruption or manipulation.

Furthermore, the system is designed to be resilient to "model bias drift." LLMs, if not properly governed, can develop biases toward the most common clinical presentations, potentially ignoring the rare but significant disparities faced by minority populations. Our infrastructure includes a continuous "adversarial fairness" loop, where the system proactively tries to find scenarios where the model might fail to identify a disparity. By intentionally challenging the model with synthetic, edge-case clinical scenarios, we harden the reasoning engine against the biases that often plague centralized healthcare AI. This proactive approach to robustness ensures that the infrastructure remains a reliable tool for public health officials even as the clinical landscape evolves.

6. Environmental Sustainability and Distributed Medical AI

The deployment of large-scale AI infrastructures carries a significant environmental footprint, often overlooked in clinical research. Centralized GPU clusters require massive amounts of energy for both computation and cooling. By distributing the LLM workload to the edge, our framework significantly reduces the energy overhead associated with massive data center operations. Edge nodes, often running on specialized, low-power AI accelerators, can perform inference with a fraction of the energy required by a general-purpose cloud server. This "compute-efficiency" is a primary structural goal, aligning health equity goals with global environmental sustainability targets.

Sustainability also extends to the lifecycle of the infrastructure. Healthcare providers, particularly those in under-resourced regions, cannot afford to constantly upgrade their hardware to keep pace with the latest AI models. Our framework addresses this through "hardware-agnostic distillation." The global cloud model is continuously distilled into multiple versions of varying complexity, allowing the system to deploy a model that matches the specific hardware capabilities of a local clinic. This prevents the "forced obsolescence" of medical infrastructure and ensures that even older compute nodes can remain part of the equity-monitoring network.

Furthermore, we advocate for "carbon-aware orchestration." The global cloud layer can schedule intensive model-retraining or large-scale synthesis tasks to occur during periods of high renewable energy availability in its local grid. This approach treats the infrastructure as a dynamic part of the energy ecosystem rather than a static consumer. By integrating sustainability into the core architectural logic, we ensure that the pursuit of health equity does not come at the cost of the environmental health of the communities we aim to serve. This holistic view of sustainability is essential for the long-term viability of distributed healthcare AI.

7. Governance, Fairness, and Algorithmic Accountability

Governance in a decentralized healthcare infrastructure is a multifaceted challenge that transcends traditional institutional boundaries. When equity assessments are automated via

LLMs, the question of "who is responsible" becomes paramount. We propose a governance framework based on "distributed accountability," where the clinical providers, system engineers, and public health officials share responsibility for the system's outputs. This is operationalized through a blockchain-based audit trail that records every model update and global synthesis event. While the patient data remains private, the logic of the disparity identification is transparent and verifiable, allowing for rigorous third-party auditing.

Fairness in this infrastructure is defined not just by the absence of bias but by the active pursuit of inclusivity. Traditional health equity research is often reactive, identifying care gaps after they have caused significant harm. Our distributed LLM framework allows for "proactive fairness," identifying emerging disparities in real-time. To ensure this is done ethically, the system incorporates a "socio-technical steering committee"—a diverse group of clinicians, community advocates, and ethicists who define the fairness constraints that the LLM must adhere to. These constraints are then translated into technical "guardrails" within the model's reasoning engine, ensuring that the pursuit of equity is grounded in community values.

Policy implications are equally profound. The use of distributed AI to identify care gaps requires new regulatory frameworks that recognize hardware-verified privacy as a sufficient substitute for traditional data-sharing agreements. We argue for a "safe harbor" policy for regional clinics that participate in equity-monitoring networks, protecting them from liability if the system identifies a disparity, provided they are taking steps to mitigate it. This shifts the focus of healthcare regulation from the punishment of failure to the collaborative identification and resolution of systemic inequities. Governance, therefore, becomes an enabling force for health equity rather than a bureaucratic hurdle.

8. Socio-Technical Implications and the Future of Regional Care

The integration of a distributed cloud-edge infrastructure for health equity marks a shift in the social contract of healthcare. It moves the responsibility for identifying disparities from centralized government agencies to a collaborative network of local providers. This empowerment of regional clinics is a significant socio-technical transformation. By providing under-resourced clinics with the same high-level analytical tools as major medical centers, the infrastructure helps to level the "digital health divide." This democratization of clinical intelligence is essential for building a truly equitable healthcare system where the quality of care is not determined by the wealth of the institution.

However, the future of regional care also depends on the "human-AI symbiosis." The LLM is not intended to replace the clinical judgment of regional physicians but to act as a "contextual assistant" that highlights care gaps they may have missed. The success of the system depends on the willingness of clinicians to engage with the AI's findings. This requires an investment in digital literacy and a clinical culture that values data-driven equity assessments. As regional clinics become active nodes in a global intelligence network, the role of the healthcare provider evolves to include a greater emphasis on population health and systemic advocacy.

Looking forward, we envision a "Global Health Equity Mesh"—a resilient, decentralized network of thousands of clinical nodes, all working together to identify and eliminate healthcare disparities in real-time. This mesh would not only track disease and care gaps but also facilitate the rapid sharing of "best practices" for equity. If a clinic in one region successfully mitigates a disparity, the system can synthesize the narrative of their success and share it as a "policy recommendation" with other clinics facing similar challenges. This collective intelligence is the ultimate goal of the distributed infrastructure, providing the structural foundation for a world where healthcare is a universal right, delivered equitably to all.

9. Conclusion

This paper has proposed a distributed cloud-edge infrastructure as a scalable and privacy-preserving solution for identifying regional healthcare disparities. By leveraging the reasoning capabilities of LLMs within a tiered architectural framework, we have demonstrated how clinical intelligence can be decentralized to protect data sovereignty while still providing global insights into health equity. Our analysis of system-level trade-offs, robustness, and sustainability underscores the necessity of a socio-technical approach to healthcare AI—one that prioritizes fairness and inclusivity as much as computational performance.

The transition toward a decentralized healthcare AI ecosystem is a critical step in overcoming the structural barriers to health equity. By building privacy, sustainability, and accountability into the core of the infrastructure, we can create tools that are not only powerful but also trustworthy. The future of healthcare lies in our ability to synthesize the diverse narratives of regional clinics into a unified commitment to care for all. As we continue to scale these privacy-preserving systems, the goal remains clear: to ensure that the digital revolution in medicine serves as a bridge to equity rather than a new source of disparity.

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. Anati, I., Gueron, S., Johnson, S., & Scarlata, V. (2013). Innovative instructions and software model for isolated execution. Proceedings of the 2nd International Workshop on Hardware and Architectural Support for Security and Privacy, 10(1).

3. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Roselander, J. (2019). Towards federated learning at scale: System design. arXiv preprint arXiv:1902.01046.

4. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

5. Chen, Y., & Sun, Y. (2020). Social commerce: A systematic review and future research directions. Journal of Business Research, 111, 1-10.

6. Costan, V., & Devadas, S. (2016). Intel SGX explained. Cryptology ePrint Archive.

7. Dwork, C. (2008). Differential privacy: A survey of results. International Conference on Theory and Applications of Models of Computation, 1-19.

8. Fu, L., Chen, X., Gao, K., Huang, X., & Tong, K. (2025, October). Memory-Augmented Knowledge Fusion with Safety-Aware Decoding for Domain-Adaptive Question Answering. In 2025 6th International Conference on Machine Learning and Computer Application (ICMLCA) (pp. 1-6). IEEE.

9. Ghoshal, B., & Tucker, A. (2022). Scalable inference for deep learning in finance. Quantitative Finance, 22(10), 1845-1860.

10. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

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

12. Kaplan, J., et al. (2020). Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.

13. Kairouz, P., McMahan, H. B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A. N., ... & Zhao, S. (2021). Advances and open problems in federated learning. Foundations and Trends® in Machine Learning, 14(1-2), 1-210.

14. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50-60.

15. Lo, A. W. (2017). Adaptive Markets: Financial Evolution at the Speed of Thought. Princeton University Press.

16. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Artificial Intelligence and Statistics, 1273-1282.

17. Mo, F., Haddadi, H., Katiyar, K., Ansari, R., & Chuah, C. N. (2021). PPFL: Privacy-preserving federated learning with trusted execution environments. Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services, 94-108.

18. Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. 2008 IEEE Symposium on Security and Privacy, 111-125.

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

20. Shalf, J. (2020). The future of computing beyond Moore's Law. Philosophical Transactions of the Royal Society A, 378(2166).

21. Stoica, I., et al. (2017). Ray: A distributed framework for emerging AI applications. 13th USENIX Symposium on Operating Systems Design and Implementation.

22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

23. Wu, S., et al. (2023). BloombergGPT: A large language model for finance. arXiv preprint arXiv:2303.17564.

24. 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).

25. Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology (TIST), 10(2), 1-19.

26. Zaharia, M., et al. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. 9th USENIX Symposium on Networked Systems Design and Implementation.

27. Zhang, L., et al. (2021). Deep reinforcement learning for automated stock trading: An ensemble strategy. SSRN Electronic Journal.

28. Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., & Chandra, V. (2018). Federated learning with non-iid data. arXiv preprint arXiv:1806.00582.

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

30. Zhang, C., Xie, Y., Bai, H., Yu, B., Li, W., & Gao, Y. (2023). A survey on federated learning for large language models. arXiv preprint arXiv:2306.05499.

31. Wang, J., et al. (2021). A field guide to federated optimization. arXiv preprint arXiv:2107.06917.

32. Rothchild, D., et al. (2020). FetchSGD: Communication-efficient federated learning with sketching. Proceedings of the 37th International Conference on Machine Learning.