Federated Fairness-Aware Learning Framework for Predicting Antiretroviral Therapy Outcomes Across Multi-Institutional Electronic Health Records in Underserved Populations

Miguel Craig
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
contactmiguel@oregonstate.edu

Chengxiang Jiang
Department of Computer Science, University of North Texas, Denton, TX, USA.
chengxiangj@unt.edu

Abstract

Antiretroviral therapy has transformed HIV infection into a manageable chronic condition, yet treatment outcomes remain uneven across demographic and socioeconomic strata, especially in underserved communities where electronic health record data are dispersed across multiple institutions. The rising adoption of machine learning for clinical outcome prediction presents both an opportunity and a structural challenge: predictive models must learn from diverse, multi-institutional data without centralizing sensitive patient information, while simultaneously mitigating disparities that arise from historically biased datasets. This paper proposes a federated fairness-aware learning framework for predicting antiretroviral therapy outcomes across multi-site electronic health records, with explicit design attention to populations experiencing systemic marginalization. We articulate a system-level architecture that couples cross-silo federated learning with fairness constraints, enabling institutions to collaboratively train a global model without exposing individual-level records. The framework addresses longitudinal missingness, demographic skew, site-specific data heterogeneity, and algorithmic fairness through a combination of local debiasing, global fairness aggregation, and continuous fairness auditing. We examine the structural trade-offs among privacy preservation, model performance, and equity objectives, highlighting tensions between local optimization and global fairness. The discussion extends to the governance infrastructure necessary for sustaining such a learning network, including institutional agreements, computational resource allocation, and policy alignment with health equity mandates. By integrating technical design with organizational and ethical considerations, this work offers a holistic blueprint for deploying equitable predictive systems in HIV care. The framework is not merely a technical artefact but a socio-technical intervention that can reshape how health systems learn from distributed data while centering the experiences of populations historically excluded from biomedical research.

Keywords

federated learning, algorithmic fairness, antiretroviral therapy, electronic health records, health equity, privacy-preserving machine learning, underserved populations.

1. Introduction

The clinical management of HIV through antiretroviral therapy (ART) has achieved remarkable reductions in morbidity and mortality, yet the benefits of treatment remain inequitably distributed across the global population and within national health systems. Underserved populations, defined by intersecting axes of socioeconomic deprivation, racial and ethnic minoritization, geographic isolation, and inadequate access to continuous care, experience systematically lower rates of viral suppression, higher rates of treatment discontinuation, and worse long-term clinical outcomes [1], [2]. Predictive models trained on electronic health record (EHR) data offer the potential to identify individuals at risk of suboptimal ART adherence or virologic failure, thereby enabling targeted interventions that can preempt adverse events. However, the data required to build such models are often fragmented across multiple healthcare institutions that serve distinct subpopulations, and conventional centralized machine learning approaches pose intractable privacy, regulatory, and logistical challenges. The imperative to respect patient confidentiality and comply with legal frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States and the General Data Protection Regulation (GDPR) in Europe makes it difficult to aggregate sensitive clinical data from diverse sites [3].

In response to these constraints, federated learning has emerged as a promising paradigm that allows multiple organizations to collaboratively train a shared model while keeping their data locally stored and never directly exchanged [4]. In cross-silo settings, a central server coordinates the aggregation of model updates sent by participating institutions, each of which trains a local model on its own private data before transmitting only gradient or parameter updates. This architecture preserves privacy at the data level while enabling statistical power that no single institution could achieve alone. Nevertheless, federated learning in its basic form does not address the pervasive problem of algorithmic bias. When participating sites serve populations with differing demographic profiles and disparate outcome distributions, a globally aggregated model can inadvertently inherit, amplify, or mask patterns of inequity present in local data [5]. For ART outcomes, these biases can manifest as models that systematically underestimate the risk of treatment failure for one racial group while overestimating it for another, or that perform well on patients with regular clinic attendance but poorly on those with fragmented care trajectories, a group disproportionately composed of marginalized individuals.

We argue that an effective learning framework for predicting ART outcomes across multi-institutional EHRs must be simultaneously federated and fairness-aware. This dual requirement introduces structural tensions: local fairness interventions implemented without coordination can diverge, leading to unstable global performance, while globally imposed fairness constraints may conflict with local clinical realities or data distributions. Moreover, notions of fairness are not monolithic; group fairness, individual fairness, and counterfactual fairness entail different operational definitions that may be incommensurable in practice, and the choice among them carries ethical and legal implications [6]. The framework we propose addresses these challenges through a modular architecture that decouples local data preprocessing and debiasing from global model aggregation and fairness auditing, thereby enabling each institution to exercise autonomy in how it handles local biases while aligning with system-wide equity objectives. This design also attends to the temporal dimension of ART care, where longitudinal EHR data are characterized by irregular observation intervals, informative missingness, and changes in clinical guidelines, all of which are amplified when data are collected in under-resourced settings.

Beyond algorithmic design, the deployment of such a framework in real-world health systems demands robust governance structures, transparent auditing mechanisms, and sustained engagement with communities that have historically been harmed by data extraction without benefit. Accordingly, this paper integrates a system-level perspective that links the computational architecture to the institutional, policy, and ethical infrastructure required for its long-term operation. We draw on lessons from large-scale health data networks, fairness in machine learning, and HIV care continuum research to develop a comprehensive design that is at once technically rigorous and socially accountable. The remainder of this paper is organized as follows. Section 2 reviews the landscape of federated learning in healthcare and fairness in predictive modeling, with attention to the specific context of ART outcomes. Section 3 describes the proposed system architecture and its federated design principles. Section 4 details the fairness-aware learning mechanisms integrated into the framework. Section 5 examines the challenges of data heterogeneity and representational bias. Section 6 discusses infrastructure, deployment, and sustainability considerations. Section 7 explores governance, ethics, and policy implications. Section 8 concludes with a forward-looking synthesis.

2. Background and Related Work

The application of machine learning to electronic health records has generated a substantial body of work aimed at predicting a range of patient outcomes, including hospital readmission, disease progression, and medication adherence. In HIV care, predictive models have been developed to forecast viral suppression, loss to follow-up, and treatment resistance using demographic, laboratory, pharmaceutical, and encounter-level features [7]. The All of Us Research Program and similar large-scale initiatives have enabled phenotyping of treatment adherence among people living with HIV, but they also reveal persistent disparities and data gaps that conventional modeling often fails to address. Existing phenotyping methods that leverage structured EHR data to measure adherence have demonstrated that suboptimal engagement with care is not randomly distributed but patterned along racial, economic, and geographic lines, underscoring the need for fairness-conscious modeling approaches. However, most current predictive tools are developed and validated within single institutions or homogeneous datasets, limiting their generalizability and potentially reinforcing site-specific biases.

Federated learning has been increasingly explored for healthcare applications, including multi-institutional prediction of mortality, length of stay, and imaging-based diagnoses. Foundational work by McMahan and colleagues introduced the federated averaging algorithm that iteratively aggregates locally trained model parameters, and subsequent variants have incorporated differential privacy guarantees, communication compression, and personalized local models [4]. Health systems such as the eICU Collaborative Research Database and the Observational Medical Outcomes Partnership (OMOP) common data model have facilitated federated analyses across disparate EHR platforms by standardizing data representation while maintaining local custody. These efforts demonstrate the feasibility of federated architectures in medical contexts, but they primarily focus on predictive performance rather than fairness. Some recent proposals have combined federated learning with fairness constraints, for instance by incorporating regularization terms that penalize disparities in model performance across sensitive attributes, or by adopting multi-objective optimization that trades off accuracy against equity metrics [8]. These approaches, while theoretically grounded, often assume relatively balanced data across sites and neglect the unique structural challenges that

arise in networks serving highly underserved populations, where site-level sample sizes may be small, class imbalances extreme, and follow-up data incomplete.

The fairness-aware machine learning literature provides a rich taxonomy of bias mitigation strategies, broadly categorized into pre-processing, in-processing, and post-processing methods [9]. Pre-processing techniques transform training data to reduce associations between sensitive attributes and outcomes; in-processing methods incorporate fairness penalties into the objective function during model training; post-processing approaches adjust model outputs after training to satisfy specific fairness criteria. Each category entails different trade-offs with respect to model interpretability, legal defensibility, and the ability to audit outcomes over time. In the federated setting, these methods must be adapted to a distributed environment where sensitive attributes may not be shared across sites, requiring either secure aggregation of fairness statistics or proxy-based approximations that avoid direct transmission of protected demographic information. The intersection of federated learning and fairness is thus a nascent but rapidly growing area, and its application to ART outcomes in underserved populations remains largely uncharted territory.

The HIV care continuum provides a compelling use case because it encapsulates longitudinal, multi-faceted engagement with health services, from diagnosis and linkage to care, through retention and viral suppression. Disparities along this continuum are well-documented: Black and Hispanic individuals in the United States, as well as people who inject drugs, experience disproportionately high rates of disengagement despite the availability of effective therapy [1]. Predictive tools that fail to account for these structural inequities risk misallocating resources away from those who need them most, while a fairness-aware approach could direct targeted retention interventions to populations with historically poorer outcomes without stigmatizing individual patients. This dual objective, improving overall system efficiency while advancing equity, motivates the design of a learning framework that is sensitive to both federated data realities and fairness imperatives.

3. System Architecture and Federated Design Principles

The proposed framework is structured as a cross-silo federated learning system in which a central aggregation server coordinates multiple institutional clients, each representing a healthcare organization that holds longitudinal EHR data for a cohort of people living with HIV. Unlike typical federated schemes, the architecture explicitly includes components for local fairness processing, global fairness aggregation, and longitudinal data harmonization, reflecting the reality that ART outcomes unfold over many years and are subject to irregular patterns of observation. Each client site maintains a local data warehouse containing structured and unstructured clinical data, transformed into a common data model that preserves semantic interoperability while allowing site-specific extensions. The choice of a common data model, such as OMOP, is strategic: it facilitates standardized feature extraction without requiring sites to reveal raw data, and it supports distributed queries that can compute population-level summaries needed for fairness monitoring.

Privacy preservation in this architecture rests on two main pillars: data remain on local servers throughout the training process, and the information transmitted to the aggregator consists solely of model parameter updates, optionally perturbed by differentially private noise. Local training can be performed on de-identified data, and the aggregator never accesses patient-level health information. To prevent membership inference or gradient leakage attacks, each site can clip gradients and add calibrated Gaussian noise before transmission, following the principles of differential privacy [3]. The trade-off between privacy budget and model utility

must be carefully managed, as stricter privacy guarantees may degrade predictive accuracy, particularly for small subgroups that are already underrepresented. The framework thus allows each institution to configure its privacy parameters based on its risk assessment and regulatory environment, while the aggregation server enforces a minimum privacy threshold to protect network-wide trust.

A distinguishing feature of the architecture is the separation of local model training into two interrelated tracks: a prediction track and a fairness track. The prediction track trains a base model, such as a recurrent neural network or a transformer-based architecture, capable of handling irregularly sampled longitudinal EHR sequences to forecast individual-level ART outcomes, including viral suppression status, retention in care, or adverse drug events. The fairness track operates in parallel, using only local data to estimate fairness metrics such as equalized odds or demographic parity with respect to predefined sensitive attributes, which are retained locally but never shared. During each communication round, clients transmit not only their model updates but also aggregated fairness statistics computed in a secure and anonymized manner. The global aggregator then combines prediction updates using federated averaging, while also computing a global fairness loss that is a weighted combination of local fairness metrics. This dual-stream design ensures that fairness considerations are embedded in the optimization process rather than applied as an afterthought.

The communication protocol is asynchronous to accommodate sites with varying computational capacities and patient volumes, a critical consideration when including safety-net hospitals and community health centers that often lack high-performance computing infrastructure. Asynchronous updates allow the central server to incorporate contributions from slower sites without blocking faster ones, though they introduce challenges in model staleness and convergence guarantees. To mitigate these challenges, the framework employs a versioned model repository and a staleness-weighted aggregation scheme that adjusts the influence of each update based on the number of rounds elapsed since the site last participated [10]. This design ensures that clinics serving marginalized populations, which may operate with limited IT staff and intermittent connectivity, remain meaningful contributors rather than being structurally excluded from the collaborative learning process.

Longitudinal data harmonization is another core architectural component. Because different sites may collect laboratory tests at different frequencies, use different coding systems for medications, and have varying follow-up windows, the framework includes a local preprocessing module that transforms raw EHR sequences into a standardized, time-aligned representation using clinically informed time windows and indicator-based alignment. Missing data are handled not through simple imputation but through learned embeddings that treat absence as informative, capturing patterns such as clinic disengagement that are themselves predictive of ART outcomes and correlated with social determinants of health. This design choice is grounded in the observation that missingness in underserved populations is often structurally driven and should not be erased by naive imputation that assumes data are missing at random.

The aggregator also houses a global fairness auditor that operates independently of the model training loop, periodically evaluating the fairness of the aggregated global model on held-out local evaluation sets, if sites consent to share evaluation metrics. The auditor generates disparity reports that are accessible to a multi-institutional governance board, which can then decide whether to adjust fairness weights, rebalance site contributions, or pause deployment until identified harms are addressed. By separating auditing from training, the architecture

introduces a layer of institutional accountability that provides recourse for affected communities and avoids conflating optimization with oversight.

4. Fairness-Aware Learning Mechanisms

Fairness in predictive modeling for ART outcomes cannot be reduced to a single mathematical constraint; it requires a multifaceted strategy that addresses representational harm, allocative harm, and temporal bias. Representational harm occurs when the model's feature space inadequately captures the lived experiences of marginalized groups, for instance by relying on clinical encounters that are infrequent for people facing transportation barriers, or by using laboratory values that are affected by structural differences in reference ranges. Allocative harm arises when model predictions lead to differential access to care resources, such as intensified case management or adherence counseling, in ways that systematically disadvantage certain groups. Temporal bias is introduced when longitudinal EHR data are unevenly collected across groups, causing the model to learn erroneous relationships between observation density and risk.

The framework incorporates three complementary fairness mechanisms: local fairness regularization, global fairness aggregation, and post-aggregation recalibration. Local fairness regularization operates during local model training by adding a penalty term to the loss function that discourages disparate impact across protected groups, using a smoothed approximation of group fairness metrics that is differentiable and amenable to gradient-based optimization. Crucially, this regularization is configured locally to reflect each site's demographic makeup and institutional priorities; a clinic serving a predominantly African American population may focus on minimizing racial disparities in viral suppression prediction, while another site with large populations of Spanish-speaking patients may target language-based fairness. This pluralistic approach acknowledges that fairness is not a universal template but must be contextually grounded.

Global fairness aggregation synthesizes these local fairness objectives into a system-wide signal without requiring sites to reveal their sensitive attribute distributions. The aggregator receives local statistics about fairness violations, expressed as differences in true positive rates or false positive rates across groups, and computes a weighted average where the weights can be set by a governance board to prioritize sites with larger underserved populations or historically greater disparities. Alternatively, the aggregation can employ a minimax objective that minimizes the worst-case fairness violation across all participating sites, thereby preventing any single site from being egregiously disadvantaged by the global model. This approach aligns with the philosophical principle of prioritarianism, giving greater weight to improving outcomes at the lower end of the equity distribution.

Post-aggregation recalibration addresses the fact that even a fairness-regularized global model may still exhibit biases when applied to local populations that differ from the aggregate in important ways. After receiving the global model, each site can perform a local recalibration using isotonic regression or Platt scaling, where the calibration is stratified by sensitive groups to achieve within-group calibration without suppressing legitimate between-group differences that reflect differential disease burden. This step ensures that predicted probabilities reflect actual risk within each demographic stratum, which is essential for clinical decision-making. For instance, a threshold-based alert system for ART adherence support will perform equitably only if the predicted risk scores are comparably calibrated for Black and white patients at each site.

The framework further includes a dynamic fairness monitoring component that tracks fairness metrics over time as the model is updated and as population demographics evolve. Temporal concept drift, where the relationship between features and outcomes changes, can erode fairness even if the model was initially well-calibrated. Drift is especially relevant in HIV care due to changes in treatment guidelines, introduction of long-acting injectable ART formulations, and demographic shifts from migration or policy changes. The monitoring module compares current fairness metrics against historical baselines using statistical process control and raises alerts when deviations exceed prespecified thresholds. These alerts trigger a review process involving the governance board and affected sites, reinforcing the principle that fairness is not a one-time design goal but an ongoing operational commitment.

To address the specific challenge of intersectional fairness, where individuals belong to multiple protected groups simultaneously, the framework encourages sites to compute fairness metrics for intersectional subgroups where sample sizes permit and to report these in de-identified aggregate form to the global auditor. While intersectional analysis exacerbates the small-sample problem, particularly in smaller clinics, the framework adopts a hierarchical fairness reporting structure that starts with single-axis metrics and deepens to intersectional analysis as data accumulate over time and across federated rounds. This layered approach avoids the paralysis that can result from demanding granular intersectional data that local institutions may not be able to reliably provide, while still creating a pathway toward more comprehensive equity evaluation.

5. Data Heterogeneity and Representational Challenges

Data heterogeneity across participating institutions is both a resource and a barrier for fairness-aware federated learning. Heterogeneity arises from differences in EHR platforms, clinical coding practices, patient demographics, treatment protocols, and local care patterns, all of which are amplified when the network includes federally qualified health centers, Ryan White HIV/AIDS Program clinics, and academic medical centers that serve vastly different patient populations. This diversity enriches the global model's exposure to varied clinical presentations and social contexts, potentially improving its robustness and generalizability. However, it also introduces statistical challenges, including non-identically distributed data across sites, label skew where viral suppression rates can vary from 50% to over 90%, and feature skew where certain laboratory tests or social history variables are recorded only at a subset of sites. Standard federated averaging assumes that local optima are not radically divergent, an assumption that breaks down under severe feature and label distribution shifts, leading to unstable convergence or a global model that performs well only on the majority demographic profile.

The framework addresses these heterogeneity challenges through a combination of local model personalization, clustered federated learning, and reweighting strategies. Personalization allows each site to fine-tune the global model on its local data after aggregation, producing a set of related but site-adapted models that capture local idiosyncrasies while sharing a common representational backbone. This approach acknowledges that a single universal model may not serve all populations equitably, and that allowing for controlled local adaptation can preserve fairness and accuracy simultaneously. However, excessive personalization risks fragmenting the knowledge base and reducing the benefits of collaboration, so the framework imposes regularization that penalizes large deviations from the global model unless justified by demonstrable local performance improvements.

Clustered federated learning offers an intermediate solution by grouping sites with similar data distributions and fairness profiles into sub-networks that coordinate more tightly among themselves before contributing to the global model. Site clustering can be based on demographic profiles, care setting types, or observed outcome distributions, and the aggregator can maintain multiple cluster-level models that compete or merge during the federated process. This architecture respects the reality that the optimal fairness strategy for urban safety-net clinics may differ from that for suburban specialty practices, while still enabling cross-cluster knowledge transfer through the global aggregation layer. The computational overhead of maintaining multiple clusters is managed by the central server using dynamic resource allocation, which can scale up cluster processing during periods of low network load.

Representational challenges extend beyond standard statistical heterogeneity to encompass the fundamental question of whose data are recorded in EHR systems and how those recordings encode systemic biases. Underserved populations are more likely to experience fragmented care across multiple institutions, leading to incomplete or duplicated records that distort the longitudinal picture. Social determinants of health, such as housing instability, food insecurity, and incarceration history, are routinely undercoded in structured fields, yet they profoundly influence ART adherence and outcomes. The framework's preprocessing module incorporates natural language processing on unstructured clinical notes, with strict privacy controls, to extract additional context about social adversity, but it does so only at institutions that have obtained appropriate consents and institutional review board approvals for such analyses. This opt-in approach respects local ethical standards while creating a richer, though heterogeneous, feature landscape.

The tension between standardizing data for algorithmic tractability and retaining the contextual richness that reflects social reality is a persistent design dilemma. Over-standardization can erase the very signals that distinguish underserved populations, while under-standardization impedes interoperability. The framework navigates this dilemma through a flexible data schema that mandates a minimal set of foundational clinical variables for all sites, including ART regimens, CD4 counts, viral load measurements, and visit dates, while allowing each site to contribute additional structured and unstructured features within a common extension mechanism. This approach ensures a basic level of inter-site comparability without stifling the contextual detail that is essential for fair and accurate predictions. The global aggregator is designed to accommodate models with heterogeneous input spaces by employing feature-aligned aggregation that averages parameters only for shared model components while leaving site-specific extensions unchanged during global updates.

6. Infrastructure, Deployment, and Sustainability

Deploying a fairness-aware federated learning framework across multiple healthcare institutions demands a robust technical infrastructure as well as sustainable organizational mechanisms that extend far beyond a single research project. The computational infrastructure must support secure local training at each site, encrypted communication channels, a reliable central aggregation server, and a distributed monitoring system. In practice, this infrastructure must be compatible with the existing IT environments of diverse clinical organizations, many of which operate on legacy EHR systems with limited application programming interface support. Containerization technologies, such as Docker, combined with federated learning platforms that provide pre-built connectors to common EHR databases, can significantly reduce the site-specific engineering burden, enabling clinics with small IT teams to participate

without custom development. Still, the initial effort of mapping local data to a common data model and validating feature definitions requires dedicated data steward time, which must be supported by grants or institutional commitments.

The central aggregation server is ideally hosted in a neutral, trusted environment such as a university-based research computing center or a federally supported health data enclave, rather than at any single participating institution, to avoid perceptions of data power imbalances. The server infrastructure must be designed for high availability and resilience, with automated failover and redundant storage, because interruptions in aggregation can stall the entire learning cycle and erode trust among participants. The framework includes a version control system for models and fairness configurations, allowing sites to roll back to previous stable versions if a new aggregated model introduces regressions in local fairness or performance. This operational transparency is crucial for maintaining the confidence of clinical stakeholders who are ultimately responsible for patient care decisions influenced by the model.

Sustainability of such a network requires moving beyond grant-funded pilot phases toward a model of shared ownership and long-term institutional support. We envision a federated consortium governed by a multi-stakeholder board that includes representatives from participating clinics, patient communities, public health agencies, and ethics scholars. The board would oversee resource allocation, establish data contribution and usage policies, and adjudicate disputes over fairness violations. Financial sustainability could be achieved through a combination of membership fees scaled to organizational size, public health funding tied to health equity outcomes, and value-based reimbursement arrangements where payers contribute to the infrastructure in exchange for improved population health metrics. The model must avoid creating a two-tiered system where only well-resourced academic medical centers can afford participation, which would undermine the core equity mission. Accordingly, the framework recommends a sliding-scale fee structure and in-kind contributions of computing resources by larger institutions on behalf of smaller partners.

Ongoing model maintenance in a federated system is more complex than in centralized ML pipelines because updates must be carefully orchestrated across sites with differing operational calendars and compliance requirements. The framework adopts a quarterly retraining cycle as a default, with emergency update provisions if significant performance degradation is detected, such as after a major change in treatment guidelines or during a public health crisis. Each retraining cycle includes a fairness impact assessment conducted by the global auditor, whose report is shared with the governance board before the new model is deployed. Sites may then choose to adopt the updated model or continue with the previous version while participating in further refinement. This opt-in deployment model respects institutional autonomy and mirrors the clinical governance of guideline adoption in decentralized health systems.

Workforce capacity building is an often-overlooked infrastructure need. Community health centers that serve underserved populations frequently lack data scientists and machine learning engineers who can manage local training pipelines or interpret fairness reports. The framework therefore incorporates a capacity-building layer that provides training materials, open-source toolkits, and a shared help desk, supported by the consortium’s larger members. This layer is not merely a gesture but a structural investment needed to ensure that all sites can meaningfully participate in both the technical and governance aspects of the network. Without it, federated learning can reinforce existing inequities by concentrating epistemic

authority and technical control in elite institutions, exactly the dynamic that fairness-aware design seeks to counteract.

7. Governance, Ethics, and Policy Implications

The introduction of a fairness-aware federated learning framework into HIV care infrastructure raises profound governance and ethical questions that extend well beyond technical architecture. First among these is the question of who defines fairness for a given clinical context and through what processes. The framework resists the technocratic tendency to delegate fairness definitions to model developers alone, proposing instead a participatory governance process in which patient representatives, community health workers, and civil society organizations contribute to the specification of fairness metrics and the weighting of competing equity objectives. This participatory approach is grounded in the recognition that historically marginalized communities have been harmed by medical technologies designed without their input, and that trust in predictive systems for HIV care can only be rebuilt through sustained, transparent engagement.

The legal landscape governing multi-institutional health data sharing is complex, and federated architectures do not fully resolve all regulatory exposures. While data remain local, the transmission of model updates can still be subject to legal agreements governing protected health information, especially if gradients could be reverse-engineered under concerted attack. The framework therefore embeds a legal governance layer that standardizes participation agreements using templates that specify allowable data uses, privacy safeguards, intellectual property arrangements, and liability allocation in the event of adverse patient outcomes attributable, even partially, to model predictions. These agreements must be crafted with awareness of the power differentials between large academic institutions and community-based clinics, ensuring that smaller partners retain meaningful control over how their data contributions are used and are not coerced into arrangements that primarily benefit more powerful actors.

Policy alignment with broader health equity initiatives is essential for the framework to achieve sustained impact rather than remaining a boutique research artifact. In the United States, the Ending the HIV Epidemic initiative and the expansion of the Ryan White HIV/AIDS Program provide policy hooks that could support the deployment of fairness-aware predictive tools, especially if those tools can demonstrate measurable reductions in disparities. Regulatory frameworks should also evolve to recognize the unique risk-benefit calculus of federated models that explicitly target equity, perhaps through expedited safe harbor provisions for health systems that implement audited fairness practices. Conversely, the framework must guard against the risk that predictive models, even fairness-aware ones, become instruments of surveillance or conditionality that penalize patients rather than support them, a concern that has deep resonance in communities with histories of coercive public health interventions.

Ethically, the framework commits to what might be termed restorative fairness, a principle that the distribution of benefits from the learning system should preferentially flow to the communities whose data made the model possible. This implies that if the model enables more efficient allocation of adherence support interventions, those interventions should be made available first to the clinics and populations that contributed data and that bear the highest burden of HIV-related disparities. Operationalizing restorative fairness requires tracking the flow of benefits post-deployment and reporting transparently to the governance board on whether the model’s use is closing or widening equity gaps. This accountability

mechanism transforms the framework from a passive technical service into an active participant in the health equity ecosystem.

The international dimension of ART outcomes is also relevant, though the framework described here is primarily situated within high-resource health systems with robust EHR infrastructure. The principles of federated fairness-aware learning are, however, adaptable to low- and middle-income settings where mobile health data and simpler clinic registries serve as the primary data sources. Extending the framework to global contexts would require re-engineering the data harmonization layer for lower digital maturity environments and engaging with community health workers, who often bridge formal and informal care systems. While such an extension is beyond the present scope, the system architecture is designed with modularity that could accommodate these adaptations, and the governance principles are intentionally broad enough to support transnational equity coalitions.

8. Conclusion

This paper has advanced a federated fairness-aware learning framework tailored to the prediction of antiretroviral therapy outcomes across multi-institutional electronic health records, with a deliberate focus on populations that experience persistent health disparities. By interlacing cross-silo federated learning with local and global fairness mechanisms, the framework addresses the dual imperatives of privacy preservation and equitable predictive performance in a domain where both are ethically non-negotiable. The system architecture weaves together technical components, from differentially private gradient transmission to asynchronous aggregation and longitudinal missingness handling, with organizational components, including a neutral aggregator, a capacity-building layer, and a multi-stakeholder governance board. This integrated design reflects our conviction that equitable AI in healthcare cannot be achieved by algorithmic fixes alone; it requires structural reforms in how health data are governed, who participates in defining fairness, and how the benefits of learning systems are distributed.

The challenges ahead are considerable. Empirical validation of the framework in a real-world multi-site HIV network is the next essential step, one that will surface unanticipated tensions between privacy budgets and fairness metrics, between local autonomy and global coordination, and between the pace of machine learning innovation and the deliberate tempo of community-engaged governance. Such validation must be conducted not as a purely technical exercise but as a collaborative learning process with clinics and communities, in which the framework itself is refined based on lived experiences of deployment. The conceptual architecture presented here provides both a foundation and a roadmap for that journey. By centering the perspectives of underserved populations and by embedding fairness into the very fabric of federated learning, we can move beyond reactionary mitigation of biased models toward the proactive construction of predictive systems that advance health equity as a core output rather than an ancillary aspiration.

References

1. Cohen, M. S., Chen, Y. Q., McCauley, M., Gamble, T., Hosseinipour, M. C., Kumarasamy, N., ... & Fleming, T. R. (2011). Prevention of HIV-1 infection with early antiretroviral therapy. New England Journal of Medicine, 365(6), 493–505.

2. Millett, G. A., Peterson, J. L., Flores, S. A., Hart, T. A., Jeffries, W. L., Wilson, P. A., ... & Remis, R. S. (2012). Comparisons of disparities and risks of HIV infection in black and

other men who have sex with men in Canada, UK, and USA: A meta-analysis. The Lancet, 380(9839), 341–348.

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

4. McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (pp. 1273–1282).

5. 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.

6. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (pp. 3315–3323).

7. 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).

8. Mohri, M., Sivek, G., & Suresh, A. T. (2019). Agnostic federated learning. In Proceedings of the 36th International Conference on Machine Learning (pp. 4615–4625).

9. Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and machine learning. fairmlbook.org.

10. Xie, C., Koyejo, S., & Gupta, I. (2019). Asynchronous federated optimization. In Advances in Neural Information Processing Systems (pp. 8732–8742).

11. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.

12. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. New England Journal of Medicine, 380(14), 1347–1358.

13. Gianfrancesco, M. A., Tamang, S., Yazdany, J., & Schmajuk, G. (2018). Potential biases in machine learning algorithms using electronic health record data. JAMA Internal Medicine, 178(11), 1544–1547.

14. Wiener, R. S., Gould, M. K., Woloshin, S., Schwartz, L. M., & Clark, J. A. (2013). "The thing is, what is under my control?": Patient views on computed tomography lung cancer screening. Archives of Internal Medicine, 173(11), 1025–1026.

15. 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.

16. 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.

17. Chen, I. Y., Szolovits, P., & Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care? AMA Journal of Ethics, 21(2), 167–179.

18. 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.

19. Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. Nature Medicine, 25(1), 37–43.

20. Murphy, K., Di Ruggiero, E., Upshur, R., Willison, D. J., Malhotra, N., Cai, J. C., ... & Gibson, J. (2021). Artificial intelligence for good health: A scoping review of the ethics literature. BMC Medical Ethics, 22(1), 14.

21. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care: Addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983.