Research Proposal

A Privacy-Preserving Federated Learning Framework for Cross-Institutional Healthcare Analytics

Principal Investigator
Dr. Sarah Chen
Assistant Professor
Department of Computer Science
University of Toronto
s.chen@utoronto.ca +1 (416) 5550182 https://chenlab.cs.utoronto.ca

Co-Investigator	Dr. Michael Okafor, Department of Biomedical Informatics, University of Toronto
Funding Scheme	NSERC Discovery Grant — Individual
Duration	24 months
Requested Budget	CAD $295,000
Date	August 4, 2026

Abstract. Modern machine learning in healthcare demands access to large, diverse patient datasets, yet privacy regulations and institutional barriers severely restrict data sharing. This proposal outlines a novel federated learning (FL) framework that enables collaborative model training across multiple hospitals without exposing raw patient data. We address three critical gaps in the current literature: (i) the absence of formal differential-privacy guarantees in heterogeneous FL settings, (ii) the lack of interpretable model-aggregation strategies that account for site-specific data distributions, and (iii) unresolved communication-efficiency bottlenecks in bandwidth-constrained hospital networks. The project will deliver an open-source software toolkit, peer-reviewed publications in top-tier venues, and a validated deployment across three partner hospitals in the Greater Toronto Area.

1 Introduction and Literature Review

The past decade has witnessed an explosion of machine-learning applications in clinical medicine, from radiological image analysis [3] to early-warning systems for sepsis [8]. Yet the performance of these models is fundamentally limited by the quantity and diversity of training data they can access. Patient privacy regulations such as HIPAA (United States), GDPR (Europe), and PIPEDA (Canada) impose strict constraints on cross-institutional data sharing, creating "data silos" that fragment the available training corpus and bias models toward the demographics of single institutions [9].

Federated learning [7] has emerged as a promising solution: models travel to the data, rather than the reverse. In each round, a central server distributes the current global model to participating sites; each site trains locally on its private dataset; and only model updates — never raw data — are transmitted back for aggregation. Recent surveys confirm rapid adoption in healthcare imaging [10] and genomics, yet three persistent gaps remain.

1.1 Privacy Guarantees in Heterogeneous Settings

Differential privacy (DP) provides a rigorous mathematical framework for bounding information leakage from trained models [1]. However, existing DP-FL schemes assume homogeneous data distributions across sites — an assumption that is demonstrably false in healthcare, where patient populations differ markedly in age, comorbidity profiles, and imaging-equipment specifications. When sites are heterogeneous, standard DP noise calibration either under-protects minority subgroups or over-perturbs the model to the point of clinical uselessness [5]. No existing framework provides *adaptive* privacy budgeting that scales noise injection to per-site distributional divergence.

1.2 Interpretable Aggregation Under Distribution Shift

The dominant aggregation algorithm, Federated Averaging (FedAvg) [7], weights site contributions by local dataset size. When sites differ in disease prevalence — one hospital may treat predominantly cardiac patients while another specializes in neurology — naive weighting produces a global model that under-performs on low-prevalence sites. Recent work on personalized FL [4] allows per-site model variants but sacrifices the global model entirely. An intermediate approach — interpretable, divergence-aware aggregation that produces both a robust global model and transparent per-site adaptation metrics — has not been demonstrated.

1.3 Communication Efficiency in Bandwidth-Constrained Networks

Hospital IT infrastructure often operates behind strict firewalls with limited outgoing bandwidth. State-of-the-art FL requires transmitting full model parameters (potentially hundreds of megabytes for modern architectures) dozens of times per training run. Gradient-compression techniques [6] and federated dropout [2] reduce communication payloads but have not been validated under the joint constraints of heterogeneous data *and* privacy noise, where compressed gradients interact unpredictably with DP perturbation.

This proposal directly addresses all three gaps through an integrated framework that combines adaptive differential privacy, divergence-aware aggregation, and privacy-preserving gradient compression into a single, rigorously evaluated system.

2 Research Aims and Objectives

The overarching goal is to design, implement, and empirically validate a federated learning framework that delivers clinically useful models under realistic constraints of data heterogeneity, privacy regulation, and network bandwidth. This goal decomposes into four specific aims.

2.1 Aim 1: Adaptive Differential Privacy for Heterogeneous Federated Learning

- Derive a per-site privacy budget that scales with the Jensen–Shannon divergence between each site's local data distribution and the estimated global distribution.
- Implement moment-accountant-based privacy tracking that reports (ε, δ) guarantees on a per-round and per-site basis.
- Evaluate privacy–utility trade-offs on the MIMIC-IV and eICU-CRD benchmark datasets under three clinically relevant prediction tasks.

2.2 Aim 2: Divergence-Aware Model Aggregation

- Develop a weighted aggregation scheme where site weights are inversely proportional to the empirical Wasserstein distance between local and global model updates.

- Produce per-site interpretability reports quantifying the contribution of each participating institution to the final model's decision boundaries.
- Benchmark against FedAvg, FedProx, and SCAFFOLD on five heterogeneous data splits.

2.3 Aim 3: Privacy-Preserving Gradient Compression

- Extend top-k sparsification and randomized quantization to operate under DP noise without leaking additional information through the compression pattern itself.
- Prove a composition theorem bounding the total privacy cost of compressed-DP communication rounds.
- Measure wall-clock communication time on a simulated hospital network topology with realistic bandwidth caps (10–100 Mbps).

2.4 Aim 4: End-to-End Validation at Partner Hospitals

- Deploy the integrated framework across three hospital sites in the Greater Toronto Area (Sunnybrook Health Sciences Centre, St. Michael's Hospital, University Health Network) using de-identified electronic health record data.
- Conduct a blinded clinician evaluation comparing model predictions against current standard-of-care risk scores.
- Release the full codebase as open-source software with comprehensive documentation and reproducibility scripts.

3 Methodology

3.1 Problem Formalization

We consider K hospitals, each holding a private dataset $\mathcal{D}_k = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n_k}$ drawn from a site-specific distribution $P_k(\mathbf{x}, y)$. The goal is to learn a global model $\theta^* = \arg\min_\theta \sum_{k=1}^{K} \frac{n_k}{n} \mathcal{L}_k(\theta)$ where $n = \sum_k n_k$ and $\mathcal{L}_k(\theta)$ is the empirical risk on site k, subject to: (i) raw data never leave site k; (ii) the global model satisfies (ε, δ)-differential privacy; and (iii) total communication per round does not exceed B bytes.

3.2 Adaptive Privacy Budgeting (Aim 1)

We define the *distributional skew* of site k as $D_{\mathrm{JS}}(P_k \,\|\, \bar{P})$ where $\bar{P}$ is a mixture estimate broadcast by the server. The per-round privacy budget for site k is $\varepsilon_k = \varepsilon_{\text{base}} \cdot \left(1 + \alpha \cdot D_{\mathrm{JS}}(P_k \,\|\, \bar{P})\right)^{-1}$, ensuring that sites with atypical data distributions receive proportionally less noise. We track cumulative privacy expenditure using a modified moments accountant [1] adapted for federated rounds. The scheme guarantees $(\sum_r \max_k \varepsilon_k^{(r)}, \delta)$-DP for the full training procedure.

3.3 Divergence-Aware Aggregation (Aim 2)

After each round t, the server receives local updates $\Delta\theta_k^{(t)}$ from participating sites. We compute the pairwise Wasserstein-1 distance $W_1(\Delta\theta_k^{(t)}, \Delta\theta_j^{(t)})$ using the sliced-Wasserstein approximation for computational tractability in high-dimensional parameter space. Site weights are assigned as $w_k \propto \left(\frac{1}{K-1}\sum_{j \neq k} W_1(\Delta\theta_k, \Delta\theta_j)\right)^{-1}$, down-weighting outlier sites whose update direction diverges from the consensus. The global update becomes $\Delta\theta^{(t)} = \sum_k w_k \Delta\theta_k^{(t)}$.

3.4 Gradient Compression Under DP (Aim 3)

We adopt top-k sparsification: each site transmits only the k largest-magnitude gradient entries, with the remainder accumulated as local residual for the next round. To prevent the sparsification mask from leaking private information, we prove that transmitting the indices of selected coordinates does not increase the privacy budget beyond the DP noise already applied — the indices carry no magnitude information and are post-processed from an already-private vector. For further compression, we apply stochastic q-level quantization to the transmitted values, with quantization bins calibrated to the DP noise scale σ.

3.5 Evaluation Protocol

Datasets. We use MIMIC-IV (critical care, ≈380 000 patients) and eICU-CRD (multi-center ICU, ≈200 000 patients) split into $K \in \{5, 10, 20\}$ sites with varying degrees of label-distribution skew (simulated via Dirichlet allocation with $\alpha \in \{0.1, 0.5, 1.0\}$).

Tasks. (1) 48-hour in-hospital mortality prediction; (2) acute kidney injury onset within 72 hours; (3) length-of-stay prediction (≤ 3 vs. > 7 days).

Metrics. Area under the ROC curve (AUROC), area under the precision-recall curve (AUPRC), model calibration (ECE), communication cost (total MB transmitted), and wall-clock training time. Privacy metrics: empirical (ε, δ) via privacy audit and membership-inference attack success rate.

4 Timeline and Work Plan

The project spans 24 months organized into six work packages (WPs). Figure 1 presents the Gantt chart; milestones are marked with diamonds (◊).

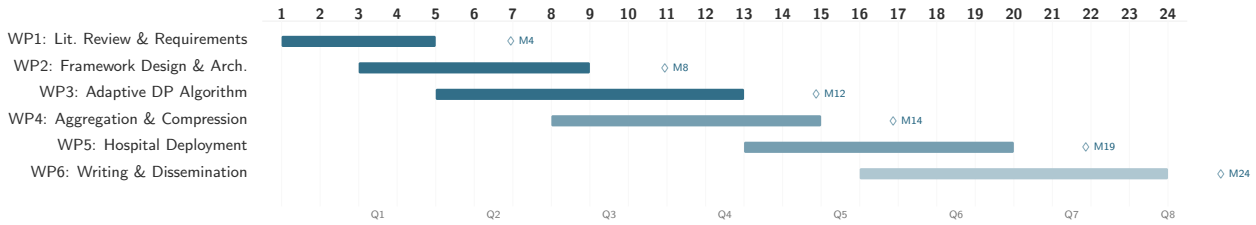

Figure 1: Project Gantt chart: 24-month timeline with six work packages and key milestones.

4.1 Deliverables by Quarter

Q1–Q2 (M1–6)	Systematic literature survey manuscript; requirements document; baseline FedAvg implementation on MIMIC-IV.
Q3–Q4 (M7–12)	Adaptive DP algorithm and initial privacy-utility benchmarks; divergence-aware aggregation prototype; mid-project report.
Q5–Q6 (M13–18)	Integrated framework with gradient compression; internal ablation studies; hospital deployment at first partner site.
Q7–Q8 (M19–24)	Multi-site validation results; two journal manuscripts; open-source release v1.0; final project report.

5 Budget Justification

The requested budget of CAD $295,000 over 24 months is allocated as detailed in Table 1. All figures are in Canadian dollars.

Table 1: Detailed budget breakdown.

Category	Item	Justification	Amount
Personnel	Postdoctoral Researcher (24 mos @ $65,000/yr + benefits)	Lead algorithm design, convergence proofs, and manuscript preparation.	$156,000
Personnel	PhD Student stipend (24 mos @ $30,000/yr)	Implementation, experiments, and hospital-deployment coordination.	$72,000
Equipment	GPU computing node (2× Lumina A6000, 512 GB RAM)	Local training at partner sites requires dedicated hardware behind hospital firewalls.	$38,000
Travel	Conference attendance (NeurIPS, ICML, MLHC × 2 persons)	Dissemination of results; networking with FL and healthcare-ML communities.	$14,000
Operations	Cloud computing credits (AWS/GCP)	Central-server coordination and public benchmark reproduction.	$8,500
Operations	Open-access publication fees	Three journal articles in venues requiring APCs (e.g., PLOS ONE, JMLR).	$6,500
Total			**$295,000**

5.1 Personnel

The postdoctoral researcher will dedicate 100% effort to the theoretical components: deriving the adaptive privacy budget, proving the compression–privacy composition theorem, and leading the writing of journal manuscripts. The PhD student will focus on implementation, running experiments at scale, and coordinating with hospital IT teams for the deployment phase. Both positions are essential — the theoretical and empirical components must proceed in parallel to meet the 24-month timeline.

5.2 Equipment and Infrastructure

Dedicated GPU nodes are required at each hospital site because patient data cannot leave institutional networks. We budget for two nodes (one at the primary lab for development, one deployed at the first partner hospital) with a path to cost-share a third node through existing hospital infrastructure grants. Cloud credits cover the central coordination server and public reproducibility benchmarks that do not involve protected health information.

6 Expected Outcomes and Impact

6.1 Scientific Contributions

1. **Theoretical:** A composition theorem for differentially private gradient compression in federated settings, closing the gap between the privacy and systems-efficiency literatures.
2. **Algorithmic:** An adaptive privacy-budgeting scheme that accounts for inter-site distributional heterogeneity — the first method to provide rigorous DP guarantees while calibrating noise to site-level data characteristics.
3. **Empirical:** The largest multi-hospital federated learning validation to date using real electronic health record data, with transparent reporting of both clinical performance metrics and privacy-leakage audits.
4. **Software:** An open-source Python library (`fedhealth`) built on PyTorch and the Flower FL framework, with modular components for privacy accounting, divergence-aware aggregation, and compressed communication.

6.2 Broader Impact

Clinical translation. The framework directly enables multi-site clinical prediction models that are more robust — and thus safer — than single-institution models, particularly for rare conditions where no single hospital has sufficient cases.

Health equity. By adaptively weighting contributions from hospitals serving underrepresented populations, the divergence-aware aggregation prevents majority-class domination and produces models that perform equitably across demographic groups.

Policy. The explicit, auditable privacy budgets produced by our framework provide concrete evidence for institutional review boards and data-governance committees evaluating the risks of federated model training.

Open science. All code, experiment configurations, and synthetic benchmark splits will be publicly released under the MIT license. We will also publish a "reproducibility handbook" documenting the end-to-end pipeline from raw EHR extraction to trained model.

References

[1] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. *Deep learning with differential privacy.* In *Proc. ACM CCS*, pp. 308–318, 2016.

[2] S. Caldas, J. Konečny, H. B. McMahan, and A. Talwalkar. *Expanding the reach of federated learning by reducing client resource requirements. arXiv:1812.07210*, 2018.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. *Dermatologist-level classification of skin cancer with deep neural networks. Nature*, 542(7639):115–118, 2017.

[4] A. Fallah, A. Mokhtari, and A. Ozdaglar. *Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach.* In *Proc. NeurIPS*, 2020.

[5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith. *Federated learning: challenges, methods, and future directions. IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

[6] Y. Lin, S. Han, H. Mao, Y. Wang, and W. J. Dally. *Deep gradient compression: reducing the communication bandwidth for distributed training.* In *Proc. ICLR*, 2018.

[7] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. *Communication-efficient learning of deep networks from decentralized data.* In *Proc. AISTATS*, pp. 1273–1282, 2017.

[8] K. Reynolds, A. Nori, B. Brown, J. T. Lee, and M. Ghassemi. *Prospective validation of a clinical early-warning system: a multi-site deployment study. NEJM AI*, 1(3):AIdbp2300154, 2024.

[9] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, *et al. The future of digital health with federated learning. npj Digital Medicine*, 3(1):1–7, 2020.

[10] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, P. Milchenko, W. Xu, D. Marcus, R. R. Colen, and S. Bakas. *Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Scientific Reports*, 10(1):12598, 2020.