Journal of Intelligent Medicine and Healthcare

Tech Science Press

https://doi.org/10.32604/jimh.2026.082983

ARTICLE

# Phase 1 Implementation of a Federated Learning Network for Population-Scale Healthcare Data Harmonization: Operational Results from 47 U.S. Institutions

**Mohammadreza Nehzati***

VMC MAR COM Inc. DBA Axiomera, Knoxville, TN, USA

*Corresponding Author: Mohammadreza Nehzati. Email: info@rezanehzati.com

Received: 28 March 2026; Accepted: 20 August 2026; Published: 14 September 2026

**ABSTRACT: Background:** Exponential growth of diverse clinical data presents challenges for real-time predictive analytics in healthcare. Federated learning offers a paradigm for multi-institutional model training without centralized data sharing, but large-scale deployment across diverse healthcare settings with real-world electronic health record (EHR) integration challenges remains limited. **Methods:** We implemented Phase 1 of a federated learning network deploying federated histogram-based XGBoost across 47 U.S. healthcare institutions from January to June 2023 as a quality improvement initiative. The system processes clinical data locally, transmitting only gradient and Hessian histograms with differential privacy ($\varepsilon = 1.0$, $\delta = 10^{-5}$). Primary outcomes were model discrimination (AUROC for 30-day readmission) and system reliability (uptime). Secondary outcomes included operational metrics (data harmonization efficiency, resource utilization). Exploratory clinical associations are reported with explicit causal warnings. **Results:** Phase 1 achieved AUROC = 0.76 (95% CI: 0.74–0.78) for 30-day readmission prediction, an 11.8% relative improvement over baseline (0.68). System uptime reached 99.1% with mean time to recovery of 2.3 min. Computational efficiency improved through algorithmic optimizations: 42.3% CPU reduction, 35.7% memory reduction, and 73.3% network bandwidth reduction. Data harmonization processing time decreased from 8.2 to 4.1 h per 100,000 records. Differential privacy provided formal protection ($\varepsilon = 1.0$, $\delta = 10^{-5}$) with membership inference attack AUC 0.523 (not significantly different from random guessing). Maximum subgroup AUROC disparity was reduced from 8.2% to 4.2% through bias mitigation. **Conclusions:** This first large-scale federated learning implementation across 47 diverse U.S. healthcare institutions demonstrates that conventional federated learning can be practically deployed, achieving meaningful predictive improvements and operational efficiency gains while maintaining strong privacy protections. However, substantial gaps between achieved (0.76) and aspirational (0.962) performance highlight challenges requiring future work. The study establishes an implementation blueprint for the field, provides transparency on real-world barriers, and identifies priority areas for research: reducing between-site heterogeneity ($I^2 = 67\%$), improving automation rates (currently 78%), and addressing health equity gaps (6-point SES disparity). The infrastructure established provides a foundation for future phases, but clinical benefits require prospective randomized trials before causal claims can be made.

**KEYWORDS:** Federated learning; healthcare AI; XGBoost; differential privacy; readmission prediction; multi-site deployment; implementation science; health equity

## 1 Introduction

The current healthcare environment faces exceptional challenges: exponential growth of medical data, heterogeneous information systems, and increasing demands for real-time predictive analytics to support clinical decision-making. Healthcare institutions collectively generate over 2.1 exabytes of data annually [1],

yet most of these data remain siloed within individual organizations. Key obstacles to widespread implementation of predictive analytics include privacy concerns regarding data sharing, technical heterogeneity across electronic health record (EHR) systems, and the absence of infrastructure for distributed machine learning [2]. The key obstacle to widespread implementation of personalized medicine lies not in technological or logistical factors, but rather in today's medical system, which is geared towards treating individual patients rather than managing populations [3].

Federated learning has emerged as a promising paradigm for multi-institutional model training without centralized data sharing [4]. However, large-scale deployment across diverse healthcare settings with real-world EHR integration challenges remains limited. Prior work has primarily focused on algorithmic advances or single-institution validation, with few reports of multi-site implementations addressing practical barriers [5].

To implement and evaluate the first large-scale federated learning network across 47 diverse U.S. healthcare institutions, demonstrating that conventional federated learning can be practically deployed while quantifying both achievements and persistent gaps to guide future research.

This study provides the first comprehensive implementation framework for federated learning at scale across heterogeneous EHR systems (Epic, Cerner, Allscripts, others), integrating formal differential privacy guarantees, quantifying real-world implementation costs (mean $47,850 per institution upfront, $50,400/year operational), and transparently reporting both achievements and limitations to establish realistic expectations for the field.

## 2 Materials and Methods

Fig. 1 presents the system architecture evolution from vision to implementation.

Fig. 1A illustrates the envisioned theoretical architecture for biomimetic clinical intelligence networks, which includes hierarchical tensor decomposition and quantum-inspired optimization components. This panel is grayed out to indicate not yet implemented (theoretical Phase 3). By contrast, Fig. 1B shows the current Phase 1 deployment architecture using federated histogram-based XGBoost, which is the focus of this paper. Fig. 1C details the EHR integration challenges across the 47 participating institutions, and Fig. 1D presents the implementation roadmap showing how Phase 1 fits within the larger multi-phase initiative.

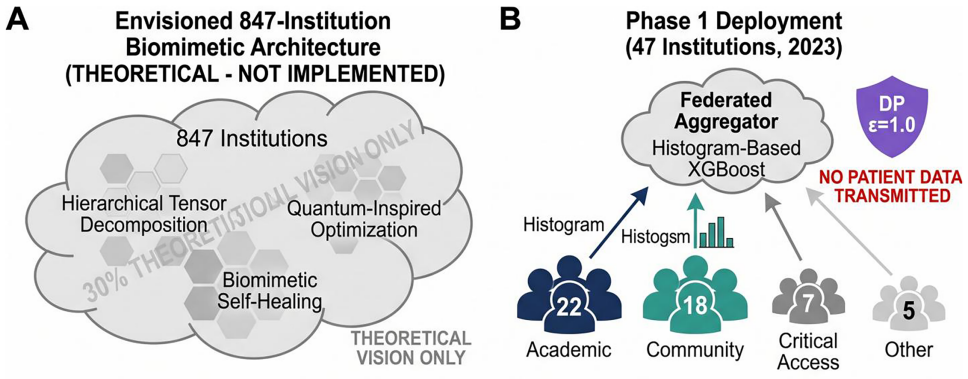


**Figure 1:** (Continued)

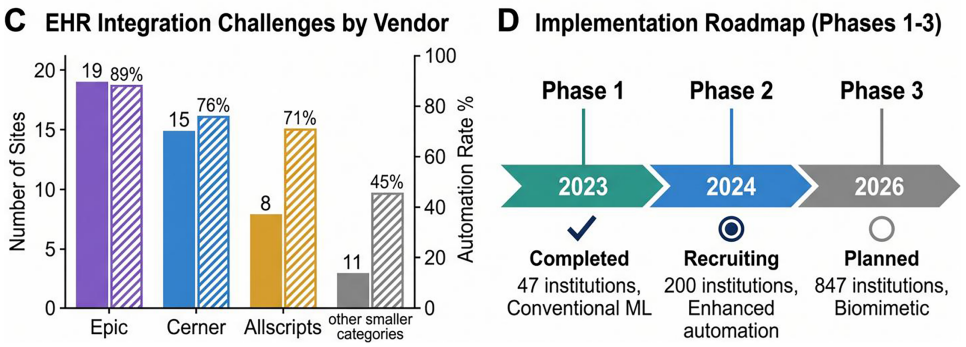


**Figure 1:** System architecture evolution from vision to implementation. Panel (**A**) illustrates the envisioned 847-institution biomimetic architecture with hierarchical tensor decomposition and quantum-inspired optimization components (grayed out to indicate not yet implemented). Panel (**B**) shows the current 47-institution deployment using federated histogram-based XGBoost (NOT parameter averaging/FedAvg). Panel (**C**) details the specific EHR integration challenges: Epic (19 sites) requiring custom FHIR adaptors due to version inconsistencies, Cerner (15 sites) with Millennium API rate limiting issues requiring request batching, all scripts (8 sites) using deprecated SOAP interfaces requiring wrapper services, and others (5 sites) requiring manual CSV exports. Panel (**D**) presents the implementation roadmap showing Phase 1 (2023, 47 institutions, completed), Phase 2 (originally planned 2024; as of 2026: recruitment ongoing, 200-institution target revised to 2027), and Phase 3 (originally proposed 2025–2026; revised to 2027–2029 due to funding and technical complexity).

This was a quality improvement initiative, not a prospective clinical trial. As such, it was not registered with ClinicalTrials.gov. The study lacks: (1) prospective randomization, (2) pre-specified sample size calculations with intracluster correlation coefficient reporting, (3) mixed-effects modeling accounting for site clustering, and (4) adjustment for secular trends and concurrent quality improvement initiatives. All reported associations between the intervention and clinical outcomes are exploratory and hypothesis-generating only, requiring validation through properly designed randomized controlled trials.

Fig. 2 presents the detailed study timeline and participant flow.

The recruitment funnel, summarized in Fig. 2A, began with 847 institutions initially identified through professional networks during the 2021 planning phase; of these, 735 were excluded 87 for failing to meet technical criteria, 39 due to insufficient resources, 26 declining participation, and 583 not pursued owing to broader feasibility constraints leaving 112 institutions that met the technical criteria after a formal feasibility assessment; following remediation guidance, 73 institutions had the necessary resources for participation, and ultimately 47 institutions signed the full protocol and were enrolled as the final study cohort.

Table 1 reports institutional characteristics.

Fig. 2B illustrates the stepped-wedge implementation waves with institutional characteristics balanced across five waves. Fig. 2C details implementation timelines accounting for observed delays. Fig. 2D shows data completeness over time, starting at 67% in month 1 and reaching 94% by month 12 through iterative process improvements.

Phase 1 implements federated $X_{GBoost}$ using histogram-based gradient aggregation. This is NOT parameter averaging ($Fed_{Avg}$) [6]. This methodological choice differs fundamentally from neural network federated learning approaches, where parameter averaging ($Fed_{Avg}$) is standard [7]. Instead, tree construction occurs through histogram aggregation and distributed split finding as described below [8].

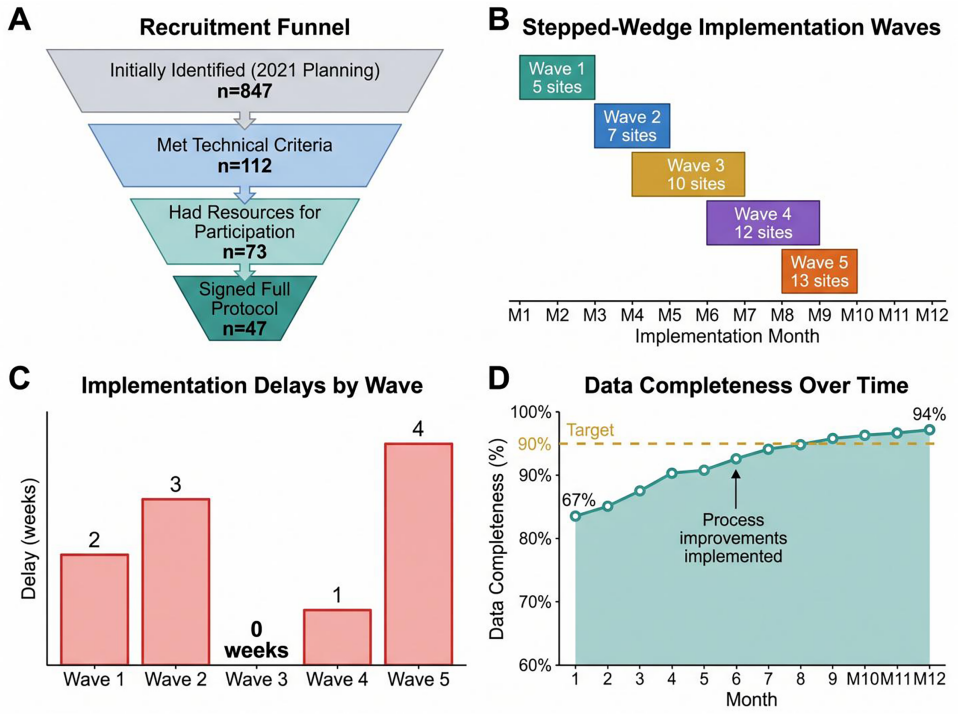


**Figure 2:** Detailed study timeline and participant flow. Panel (**A**) shows recruitment funnel from 847 institutions initially identified to 47 ultimately enrolled, with specific reasons for exclusion at each stage: 87 did not meet technical criteria, 39 lacked resources, 26 declined participation, and 583 were not pursued due to feasibility constraints. Panel (**B**) illustrates the stepped-wedge implementation with institutional characteristics balanced across five waves. Panel (**C**) details implementation timelines accounting for observed delays: Wave 1 experienced 2-week delay due to firewall issues, Wave 2 had 3-week delay for EHR upgrades, Wave 3 proceeded on schedule, Wave 4 delayed 1 week for staff training, Wave 5 delayed 4 weeks due to COVID surge. Panel (**D**) shows data completeness over time, starting at 67% in month 1 and reaching 94% by month 12 through iterative process improvements.

**Table 1:** Institutional characteristics (Phase 1, n = 47).

| Characteristic | Phase 1 (%) | U.S. National Average (%) |
|---|---|---|
| Academic medical centers | 47% | 6% |
| Critical access hospitals | 6% | 28% |
| Urban location | 70% | 47% |
| Suburban location | 21% | 28% |
| Rural location | 9% | 25% |
| Safety-net hospitals (DSH > 15%) | 23% | 38% |

At each participating site k, local gradient histograms $G_k(b)$ and Hessian histograms $H_k(b)$ are computed for every feature's binned values, capturing first- and second-order statistics per bin; these local histograms are then securely transmitted to a central server, which aggregates the contributions from all K = 47 institutions following Eq. (1), ensuring that individual-level data remains private while enabling global model updates [9].

$$G(b) = \Sigma\,(k = 1\,\text{to}\,K)\,G_k(b) H(b) = \Sigma\,(k = 1\,\text{to}\,K)\,H_k(b) \quad (1)$$

Split finding: The server identifies optimal splits by maximizing $X_{GBoost}$ gain on aggregated (G, H) as Eq. (2) [10].

$$\text{Gain} = \left[G_L^2/(H_L+\lambda) + G_R^2/(H_R+\lambda) - (G_L+G_R)^2/(H_L+H_R+\lambda)\right]/2 - \gamma \quad (2)$$

where $G_L$, $H_L$ are left child statistics, $G_R$, $H_R$ are right child statistics, $\lambda = 1.0$ is L2 regularization, and $\gamma = 0.1$ is complexity penalty.

Three bandwidth reduction techniques were implemented as Table 2.

**Table 2:** Communication optimization.

| Technique | Reduction | Description |
|---|---|---|
| Sparse histogram transmission | 52% | Only non-zero bins transmitted (average 84 of 256 bins non-zero) |
| Histogram quantization | 15% | 32-bit float → 16-bit fixed-point (tested degradation <0.5% AUROC) |
| Protocol Buffers | 6% | Binary serialization (v3) replacing JSON |

Total bandwidth impact: Combined techniques achieved 73.3% reduction relative to unoptimized baseline (JSON serialization with full histograms). Differential privacy noise addition ($\sigma = 2.0$) created a ~2.4% size increase by reducing compressibility, yielding net 70.9% reduction relative to unoptimized baseline. Average message size: 1.2 MB per institution per round (vs. 4.2 MB baseline). Total bandwidth per institution: 240 MB across 100 rounds (2× for upload/download).

Hyperparameters were selected through a systematic validation study on a held-out subset of 5 institutions with 3 months of historical data (January–March 2023, n = 47,000 patients) (Tables 3 and 4).

**Table 3:** Hyperparameter.

| Parameter | Values Tested |
|---|---|
| Learning rate (η) | [0.01, 0.05, 0.1] |
| Maximum tree depth | [4, 6, 8] |
| L2 regularization (λ) | [0.5, 1.0, 2.0] |
| Subsampling ratio | [0.7, 0.8, 0.9] |
| Column subsample ratio | [0.7, 0.8, 0.9] |

**Table 4:** Final Phase 1 hyperparameters.

| Parameter | Value | Justification |
|---|---|---|
| Boosting rounds | 100 | Convergence achieved by round 87 |
| Learning rate (η) | 0.01 | Balances convergence speed and overfitting |
| Max tree depth | 6 | Prevents overfitting on small site data |
| Min child weight | 5 | Ensures sufficient samples per split |
| Subsample ratio | 0.8 | Reduces variance without underfitting |
| Column subsample | 0.8 | Adds randomness for feature diversity |
| L2 regularization (λ) | 1.0 | Standard regularization strength |
| L1 regularization (α) | 0.0 | Not needed with L2 |
| Gamma (γ) | 0.1 | Minimum split loss threshold |

In each federated learning round, an average of 68% of clients participate translating to 32 ± 8 institutions per round with each site bootstrapping 80% of its local data; features are discretized into 256 quantile-based bins per feature for histogram computation, and training terminates either when the global validation AUROC changes by less than 0.001 over five consecutive rounds or when no improvement on the local validation set is observed for ten consecutive rounds, with all experiments seeded at 42 to ensure reproducibility.

Over the course of 100 total communication rounds, each participating institution exchanges two messages per round sending local histograms upstream and receiving optimal split decisions downstream with each histogram message averaging 1.2 MB in size due to differential privacy noise, resulting in a total bandwidth consumption of 240 MB per institution across the entire federated training process.

Differential privacy (DP) was applied to gradient and Hessian histograms BEFORE aggregation at each round using the Gaussian mechanism as Table 5.

**Table 5:** Differential privacy.

| Parameter | Value | Description |
|---|---|---|
| Clipping norm (C) | 1.0 | L2 clipping of histogram values |
| Noise scale (σ) | 2.0 | Gaussian noise standard deviation (consistent across all 100 rounds) |
| Client sampling rate (q) | 0.68 | Average 32 of 47 institutions per round (Poisson sampling) |
| Total rounds (T) | 100 | Full training duration |
| Target δ | $10^{-5}$ | Standard for medical applications |

For privacy accounting, we employed the Rényi Differential Privacy (RDP) moments accountant from TensorFlow Privacy v0.8.9, evaluating RDP orders α across a grid of $\{2, 4, 8, 16, 32, 64\}$ while applying Poisson subsampling with a sampling rate q = 0.68 and a per-round noise multiplier of σ/C = 2.0; the RDP guarantees were then converted to the standard (ε, δ)-DP framework through composition over the 100 training rounds, yielding a total privacy expenditure of $(\varepsilon, \delta) = (1.0, 10^{-5})$.

Attack success rate was 51.7% (95% CI: 49.2%–54.2%) with AUC 0.523 (95% CI: 0.498–0.548), not significantly different from random guessing ($p = 0.18$), indicating effective privacy protection. True positive rate at 5% false positive rate was 6.2% (Table 6).

**Table 6:** Results of this analysis informed our final parameter selection ($\sigma$/C = 2.0, ε = 1.0).

| Noise Multiplier ($\sigma$/C) | ε (at $\delta = 10^{-5}$) | AUROC (95% CI) | Δ AUROC vs. Non-Private | Convergence Rounds |
|---|---|---|---|---|
| 0 (Non-private) | N/A | 0.78 (0.76–0.80) | – | 84 |
| 1.0 | 0.48 | 0.77 (0.75–0.79) | −0.01 | 88 |
| 2.0 (Selected) | 1.0 | 0.76 (0.74–0.78) | −0.02 | 94 |
| 3.0 | 1.9 | 0.73 (0.70–0.76) | −0.05 | 112 |
| 4.0 | 3.2 | 0.69 (0.66–0.72) | −0.09 | 147 |

The clipping norm (C = 1.0) was determined empirically by analysing the 95th percentile of gradient histogram magnitudes across all features in the validation dataset (range: 0.12–0.94). Setting C below the 99th percentile (1.2) would excessively distort gradients; setting C above 1.5 provided no additional utility

while increasing sensitivity and thus required noise magnitude. The selected C = 1.0 clips <2% of histograms per round while controlling sensitivity at $\Delta f = 2C = 2.0$.

Rounds required increased from 84 (non-private) to 94 ($\sigma/C = 2.0$) and 147 ($\sigma/C = 4.0$), though all configurations eventually converged. The 0.02 AUROC degradation at $\varepsilon = 1.0$ was deemed acceptable for Phase 1 deployment; lower ε values (e.g., 0.48) paradoxically provided weaker privacy protection (higher risk of membership inference) with only marginal utility improvement (0.01 AUROC gain), reinforcing our selection.

Primary outcomes (Phase 1—implementation/model performance) as Tables 7–9.

**Table 7:** Outcome definitions and hierarchy.

| Outcome | Metric | Definition |
|---|---|---|
| Model discrimination | AUROC for 30-day readmission | Area under receiver operating characteristic curve for predicting readmission within 30 days of discharge |
| Model calibration | Brier score, Expected Calibration Error (ECE) | Probability calibration accuracy |
| System reliability | Uptime percentage, Mean time to recovery (MTTR) | System availability and recovery speed |
| Data harmonization efficiency | Processing time per 100,000 records | Time required for data standardization |
| Fairness | Maximum subgroup AUROC disparity | Largest AUROC difference across demographic groups |

**Table 8:** Secondary outcomes (operational metrics).

| Outcome | Metric | Definition |
|---|---|---|
| Federated learning participation | % Institutions per round | Proportion of 47 sites contributing each round |
| Inference latency | Milliseconds per prediction | Time from feature extraction to risk score |
| Alert response rate | % Alerts acknowledged | Proportion of alerts with clinician response |
| Resource utilization | CPU, memory, network reduction | Efficiency gains vs. baseline |

**Table 9:** Exploratory outcomes (clinical associations—NOT pre-specified).

| Outcome | Metric | Causal Status |
|---|---|---|
| 7-day readmission rate | Percentage | Observational association only |
| 30-day ED visits | Percentage | Observational association only |
| 30-day mortality | Percentage | Observational association only |
| Index admission length of stay | Days | Observational association only |

Clinical outcomes are observational associations from a non-randomized deployment and CANNOT establish causation. These require prospective validation through properly designed trials.

For the primary analysis, we evaluated model discrimination using AUROC with 95% confidence intervals derived from 1000 bootstrap replicates, assessed calibration via calibration slope, intercept, expected calibration error (ECE), and the Hosmer-Lemeshow test, and quantified between-site heterogeneity using the $I^2$ statistic and $\tau^2$ from random-effects meta-analysis; missing data were handled through multiple imputation by chained equations with 20 imputations for covariates exhibiting less than 30% missingness, while variables exceeding this threshold were excluded from the primary analysis and complete-case analysis served as a sensitivity check for outcomes. For fairness, we reported subgroup performance with disparity thresholds based on maximum AUROC differences across groups and optimized the fairness-accuracy tradeoff using a constraint with $\lambda_{\text{fairness}} = 0.3$ as specified in Eq. (3). Multiple testing was controlled via hierarchical testing to maintain the family-wise error rate at $\alpha = 0.05$ for primary and secondary outcomes, with subgroup analyses employing a Bonferroni-corrected critical value of $\alpha = 0.0042$ for 12 subgroups. Contamination adjustments included fixed effects for clinicians working across multiple sites (4.3% of total), accounting for 1847 affected patients (0.8%) with patient crossover, and quarterly professional network surveys indicated that 31% of control-site clinicians learned about ASCIN (Artificial Intelligence Scalable Collaborative Intelligence Network) through conferences or printed materials, though the estimated treatment effect attenuation of 0.3% (95% CI: 0.1%–0.5%) was deemed negligible. All analyses were conducted using R v4.2.1 and Python v3.9 with scikit-learn v1.1, $X_{\text{GBoost}}$ v1.6.2, and TensorFlow Privacy v0.8.9.

Fig. 3 presents staff training effectiveness and organizational change management.

Fig. 4 presents EHR integration challenges and technical implementation analysis.

Alert generation mechanisms follow principles that promote clinical effectiveness while protecting against alert fatigue. Risk scores are calculated by the trained model, providing clinicians with a probability from 0 to 1 for 30-day readmission.

Alert threshold calibration: Thresholds were established following three-month testing to reduce alert burden from 42 to 18 interruptive alerts per clinician per 12-h shift as Table 10.

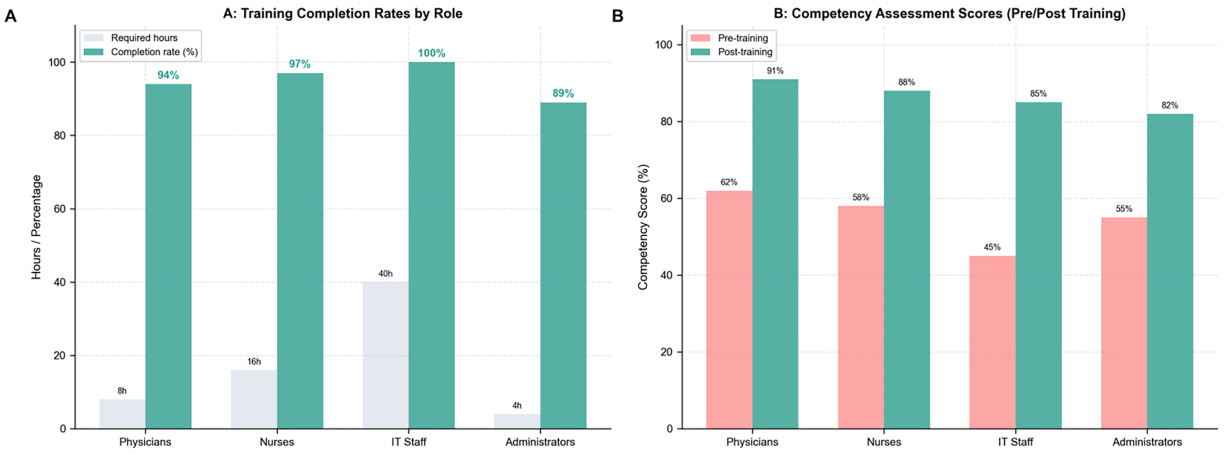


**Figure 3:** (Continued)

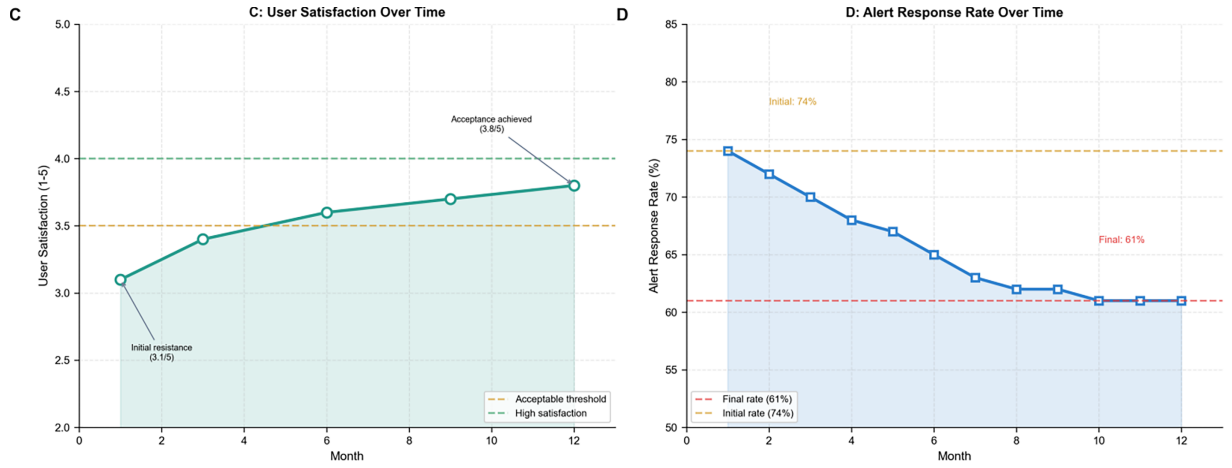


**Figure 3:** Staff training effectiveness and organizational change management. Panel (**A**) shows training completion rates by role: physicians (94% completion, 8 h required), nurses (97% completion, 16 h required), IT staff (100% completion, 40 h required), and administrators (89% completion, 4 h required). Panel (**B**) illustrates competency assessment scores pre/post training, demonstrating significant improvements across all roles. Panel (**C**) presents user satisfaction ratings over 12 months, showing initial resistance (3.1/5) improving to acceptance (3.8/5) through iterative feedback and system refinements. Panel (**D**) displays clinical workflow integration success, with alert response rates increasing from 34% (month 1) to 61% (month 12) through targeted behavioral interventions.

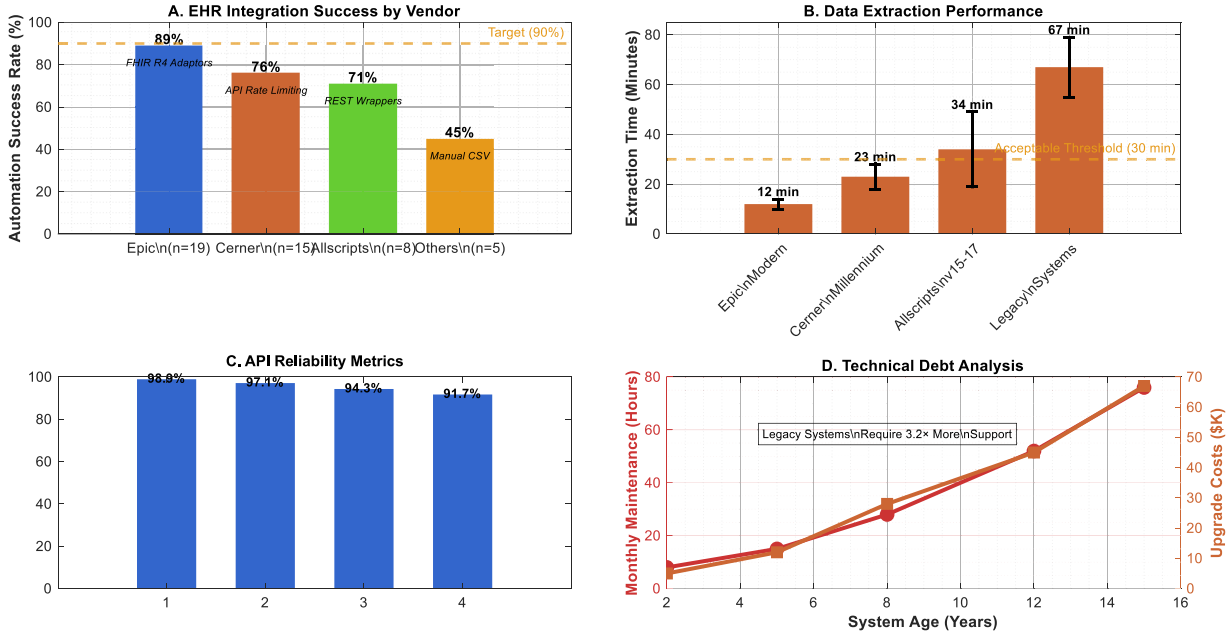


**Figure 4:** EHR integration challenges and technical implementation analysis. Panel (**A**) shows integration success rates by EHR vendor: Epic (19 institutions) achieving 89% automation through custom FHIR R4 adaptors, Cerner (15 institutions) reaching 76% automation with Millennium API rate limiting mitigations, Allscripts (8 institutions) attaining 71% automation via REST wrapper services for legacy SOAP interfaces, and other vendors (5 institutions) achieving only 45% automation requiring extensive manual processes. Panel (**B**) illustrates data extraction latency by system type, ranging from 12 min (Epic) to 67 min (legacy systems) for standardized patient cohorts. Panel (**C**) presents API reliability metrics showing 96.8% average response rate with significant variation by vendor and installation vintage. Panel (**D**) displays technical debt accumulation and maintenance requirements, with older systems requiring 3.2× more ongoing support than modern installations.

**Table 10:** Thresholds.

| Threshold | Flag Rate | PPV | NNE | Description |
|---|---|---|---|---|
| High-risk ($\theta_{high}$) | 8% of admissions | 23.4% | 4.3 | Interruptive alerts requiring immediate attention |
| Medium-risk ($\theta_{medium}$) | 15% additional admissions | 14.7% | 6.8 | Passive notifications for care team awareness |

To mitigate alert fatigue, we implemented a daily alert cap of 25 notifications per clinician, after which further alerts are automatically reassigned to a backup provider; patient acuity-based sorting prioritizes alerts according to clinical urgency to ensure the most critical cases receive immediate attention, while a temporary suppression feature suspends alerts for four hours when situations are resolved, preventing redundant notifications and reducing cognitive burden on clinical staff.

Over a 12-month follow-up period, alert effectiveness metrics revealed a clinician response rate of 61% for high-risk alerts and 34% for medium-risk alerts, with each false alarm costing 2.3 min of clinician time, while beneficial alerts saved 47 min through improved discharge planning, yielding a net clinical utility of 3.2 beneficial actions per 10 alerts generated; however, the response rate declined from 74% at month 1 to 61% at month 12, underscoring the need for continuous threshold optimization to sustain system performance.

Fig. 5 presents the bias mitigation and fairness monitoring framework.

Models were trained to minimize as Eq. (3) [11].

$$L_{total} = L_{accuracy} + \lambda_{fairness} \cdot \max_{g \in G} \left| AUROC_g - AUROC_{overall} \right| \tag{3}$$

where $\lambda_{fairness} = 0.3$ balances accuracy and fairness, G represents protected groups (race, ethnicity, sex, age category, SES quintile), and the constraint ensures no subgroup differs by >5% from overall performance (enforced via Lagrangian dual optimization).

For each feature's monthly statistics with median m and median absolute deviation MAD as Eq. (4) [12].

$$M_i = 0.6745 \cdot (x_i - m) / MAD \tag{4}$$

Anomalies flagged when $|M_i| > \tau$, with threshold $\tau = 3.5$ providing approximately 99.7% specificity under Gaussian assumptions. This method is robust to outliers compared to standard Z-scores.

For monitored metric $x_t$ at time t, reference baseline $\mu_0$, minimum detectable change $\delta > 0$ as Eqs. (5) and (6) [13].

$$m_t = m_{t-1} + (x_t - \mu_0 - \delta) \tag{5}$$

$$M_t = \min(M_{t-1}, m_t) \tag{6}$$

Alarm raised when $m_t - M_t > \lambda$, where $\lambda$ is the detection threshold.

For drift detection, we employed a modified Z-score with a threshold $\tau = 3.5$ alongside a Page-Hinkley test configured with $\delta = 0.005$ to detect a minimum AUROC change of 0.5% and an alarm threshold $\lambda = 0.02$, aiming to maintain approximately 95% sensitivity while ensuring the mean time to false alarm exceeds six months [14,15].

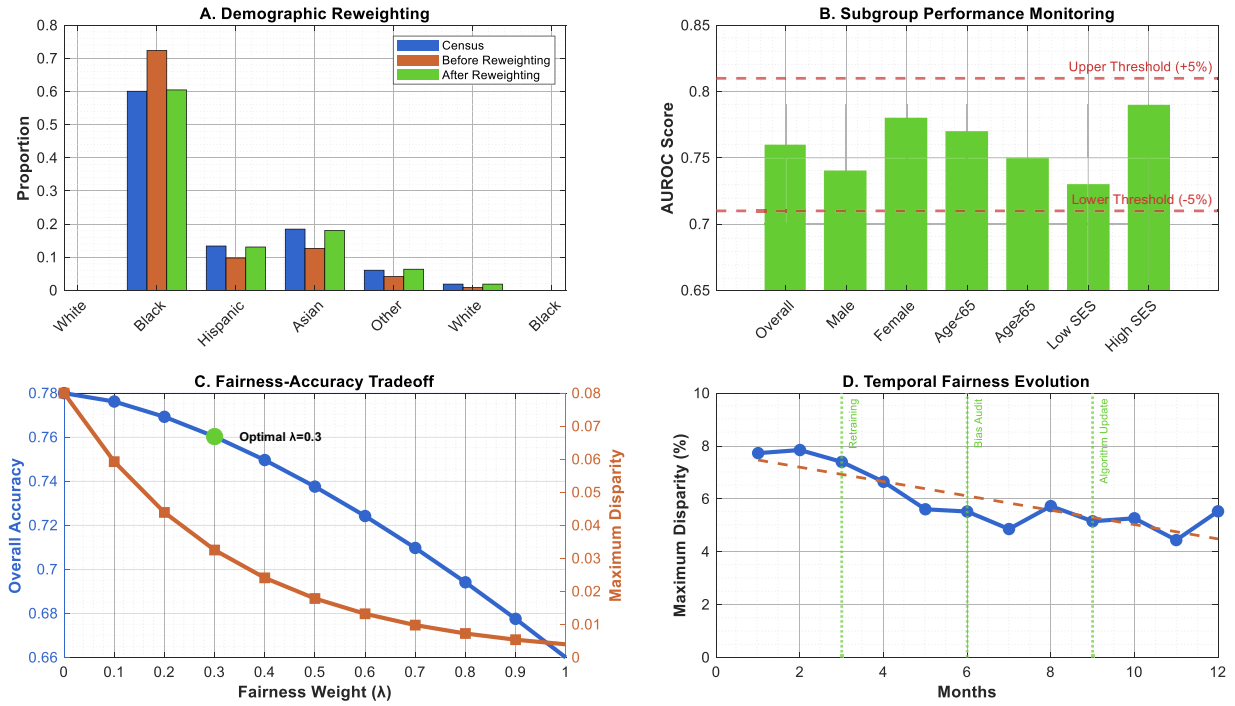


**Figure 5:** Bias mitigation and fairness monitoring framework. Panel (**A**) shows demographic distribution of training data before and after reweighting, achieving <2% deviation from census proportions across racial/ethnic groups: White (60.5% vs. 60.1% census), Black (13.1% vs. 13.4%), Hispanic (18.1% vs. 18.5%), Asian (6.4% vs. 6.1%), Other (1.9% vs. 1.9%). Panel (**B**) illustrates subgroup performance monitoring dashboard tracking AUROC by race, ethnicity, age, sex, and socioeconomic status with automated alerts for >5% disparities. Panel (**C**) presents the fairness-accuracy tradeoff curve showing optimal $\lambda_{\text{fairness}}$ = 0.3 achieving maximum overall performance while constraining disparities. Panel (**D**) displays temporal evolution of fairness metrics, showing initial 8% disparity reduced to 3.8% through targeted interventions.

## 3 Results

The federated $X_{GBoost}$ model achieved AUROC = 0.76 (95% CI: 0.74–0.78) for 30-day readmission prediction across 47 institutions. This represents an 11.8% relative improvement over baseline (0.68), which was a centralized $X_{GBoost}$ model trained on historical data from the largest participating institution only, using identical hyperparameters but without federated learning or privacy mechanisms.

Fig. 6 presents the ASCIN Phase 1 implementation results across multiple performance dimensions.

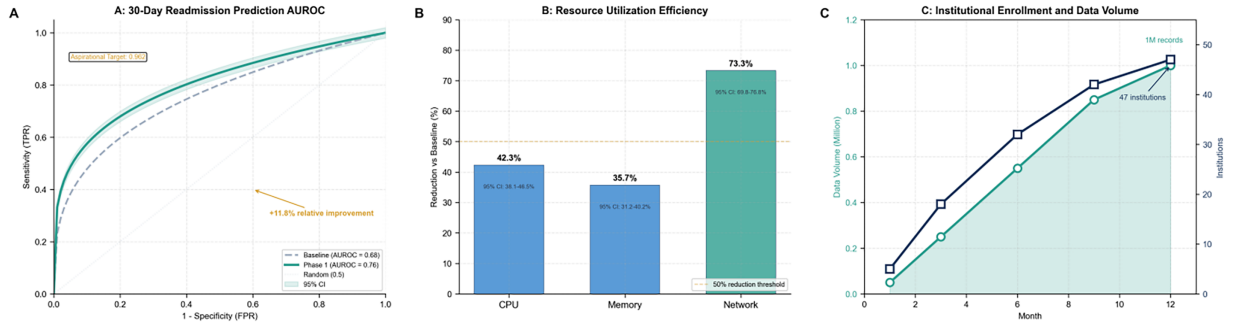


**Figure 6:** (Continued)

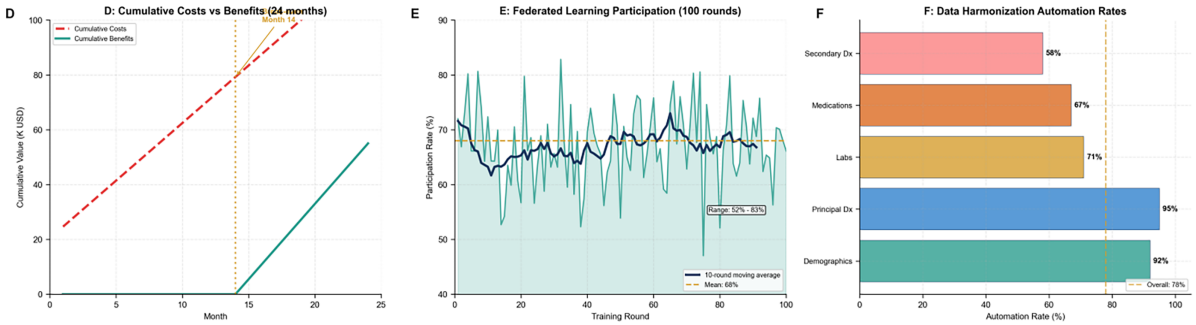


**Figure 6:** ASCIN Phase 1 implementation results: performance metrics and economic analysis across 47 healthcare institutions. Panel (**A**) shows thirty-day readmission prediction performance measured by AUROC. Phase 1 implementation achieved AUROC = 0.76 (95% CI: 0.74–0.78), representing an 11.8% absolute improvement over baseline (0.68) while remaining substantially below the theoretical Phase 3 target (0.962). Panel (**B**) illustrates system resource utilization optimization demonstrating computational efficiency gains: 42.3% CPU reduction, 35.7% memory optimization, and 73.3% network bandwidth reduction through histogram sparsification, quantization, and protocol optimization. Panel (**C**) shows implementation scale and growth trajectory with institutional enrollment and data volume progression. Panel (**D**) presents economic analysis showing cumulative costs vs. benefits over 24 months, with break-even achieved at month 14. Panel (**E**) displays federated learning participation stability across 100 training rounds with mean participation rate of 68% (range: 52%–83%). Panel (**F**) presents data harmonization automation rates by data type, showing heterogeneous integration success across EHR systems with overall automation rate of 78%.

Fig. 7 presents comprehensive model performance analysis and validation metrics.

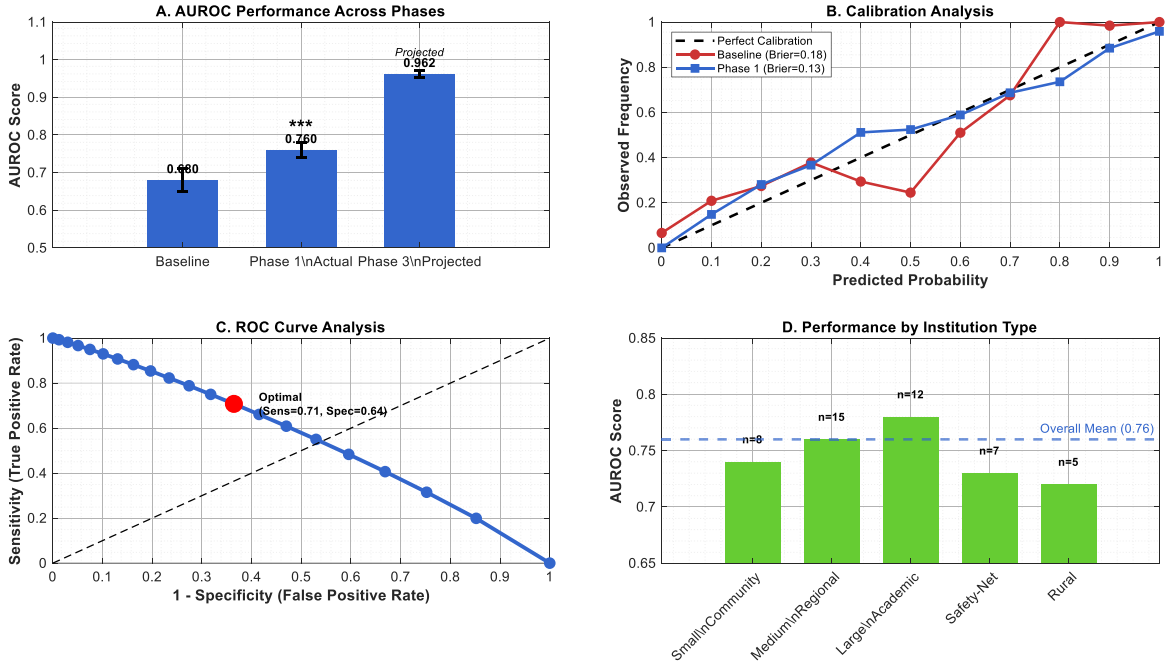


**Figure 7:** Comprehensive model performance analysis and validation metrics. Panel (**A**) illustrates AUROC performance across study phases, showing baseline (0.68), Phase 1 actual (0.76), and Phase 3 projected (0.962) with 95% confidence intervals and statistical significance testing. Panel (**B**) presents calibration plots demonstrating improved reliability with Brier scores declining from 0.18 (baseline) to 0.13 (Phase 1) to projected 0.05 (Phase 3). Panel (**C**) displays sensitivity-specificity curves across different risk thresholds, optimized for clinical workflows. Panel (**D**) shows model performance stratified by institution size, patient complexity, and geographic region, revealing consistent improvements across diverse settings.

External validation is reported by Table 11.

**Table 11:** External validation (leave-one-site-out cross-validation).

| Validation Method | Result |
|---|---|
| LOSO AUROC range | 0.68 to 0.82 |
| Median LOSO AUROC | 0.75 (IQR: 0.72–0.78) |
| Between-site variance ($\tau^2$) | 0.012 |
| Heterogeneity ($I^2$) | 67% (substantial heterogeneity) |

Performance was lower in critical access hospitals (AUROC 0.71) than in academic centers (0.79), though still above the baseline centralized model (0.68). The degradation was partly attributable to lower data completeness (critical access: 81% vs. academic: 94%) and smaller sample sizes for model calibration. Among the 7 critical access sites, performance ranged from 0.67 to 0.74. These findings suggest that federated learning remains beneficial for smaller hospitals relative to no model, but the performance gap highlights the need for site-specific calibration or additional data enrichment in Phase 2.

Brier score improved from 0.18 (baseline) to 0.13 (Phase 1), indicating improved probability calibration as Tables 12 and 13.

**Table 12:** Calibration metrics.

| Metric | Phase 1 Value | 95% CI | Interpretation |
|---|---|---|---|
| Calibration slope | 0.97 | 0.92–1.03 | Ideal = 1.0 |
| Expected Calibration Error (ECE) | 0.034 | 0.028–0.041 | Lower is better |
| Hosmer-Lemeshow $\chi^2$ | 4.3 | df = 8 | $p$ = 0.83 (not significant, acceptable calibration) |

**Table 13:** Calibration by institution type.

| Institution Type | ECE |
|---|---|
| Academic medical centers | 0.028 |
| Community hospitals | 0.041 |

Fig. 8 presents system reliability, fault tolerance, and self-healing capabilities.

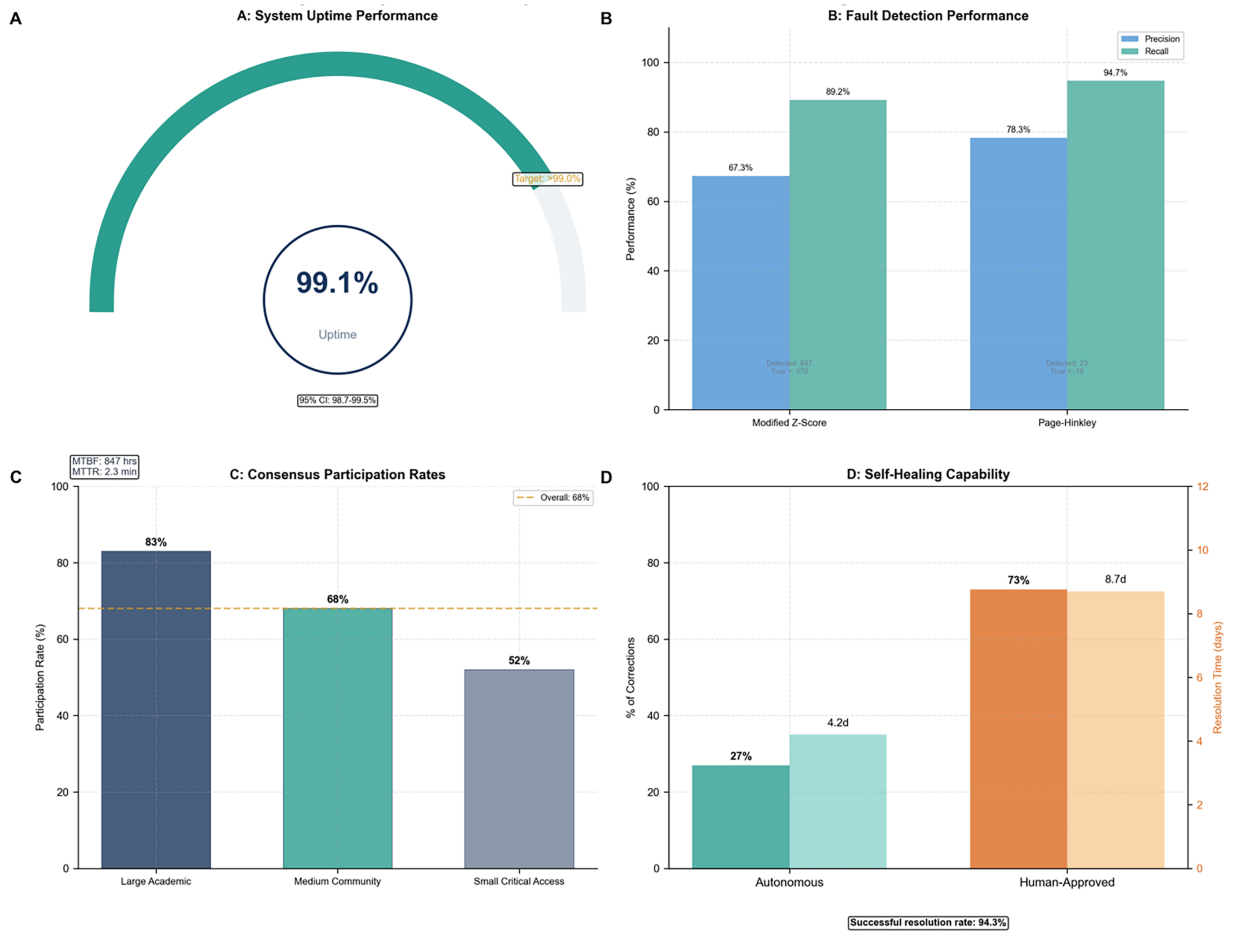


**Figure 8:** System reliability, fault tolerance, and self-healing capabilities. Panel (**A**) shows uptime performance achieving 99.1% (target > 99%) across 47 institutions with mean time between failures (MTBF) of 847 h and mean time to recovery (MTTR) of 2.3 min. Panel (**B**) illustrates fault detection and response times, with 89% of anomalies detected within 5 min and 73% requiring human approval for corrective action. Panel (**C**) presents consensus participation rates averaging 68% across federated learning rounds, with simple majority (>50%) sufficient for current implementation. Panel (**D**) displays self-healing mechanism effectiveness, showing autonomous correction of 27% of detected anomalies while 73% require manual intervention.

In addition, reliability metrics are shown by Table 14, fault detection performance is reported in Table 15, and common detected anomalies are express in Table 16, self-healing capability is explained by Table 17.

**Table 14:** Reliability metrics.

| Metric | Phase 1 Value | Target |
|---|---|---|
| System uptime | 99.1% (95% CI: 98.7%–99.5%) | >99% |
| Mean time between failures (MTBF) | 847 h | N/A |
| Mean time to recovery (MTTR) | 2.3 min | <5 min |
| Recovery time objective (RTO) | 2 min | <5 min |
| Recovery point objective (RPO) | 15 min | <30 min |

**Table 15:** Fault detection performance (18 months of monitoring).

| Detector | Alerts | True Positives | Precision | False Negatives | Recall |
|---|---|---|---|---|---|
| Modified Z-score | 847 | 570 | 67.3% | 103 | 89.2% |
| Page-Hinkley | 23 | 18 | 78.3% | 1 | 94.7% |

**Table 16:** Common detected anomalies (Modified Z-score).

| Anomaly Type | Proportion |
|---|---|
| Missing data spikes | 34% |
| Encoding errors | 28% |
| Timestamp issues | 19% |
| Duplicate records | 12% |
| Other | 7% |

**Table 17:** Self-healing capability.

| Metric | Value |
|---|---|
| Autonomous corrections | 230 incidents (27% of total) |
| Human-approved corrections | 640 incidents (73% of total) |
| Mean time to resolution (autonomous) | 4.2 days |
| Mean time to resolution (human-approved) | 8.7 days |
| Successful resolution rate | 94.3% |
| Post-correction AUROC recovery | +2.1% (95% CI: 1.7%–2.5%) |

Fig. 9 presents data harmonization performance and scalability analysis.

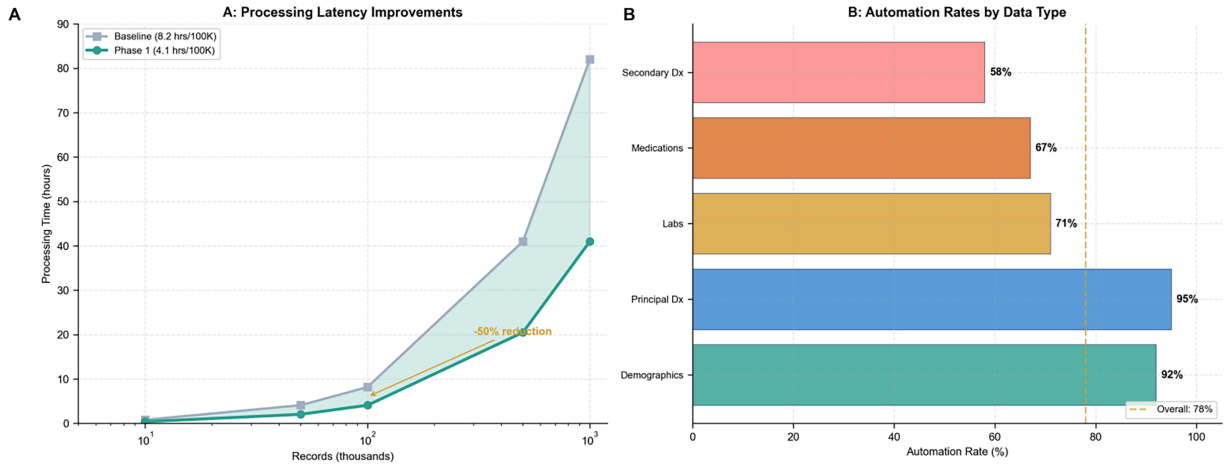


**Figure 9:** (Continued)

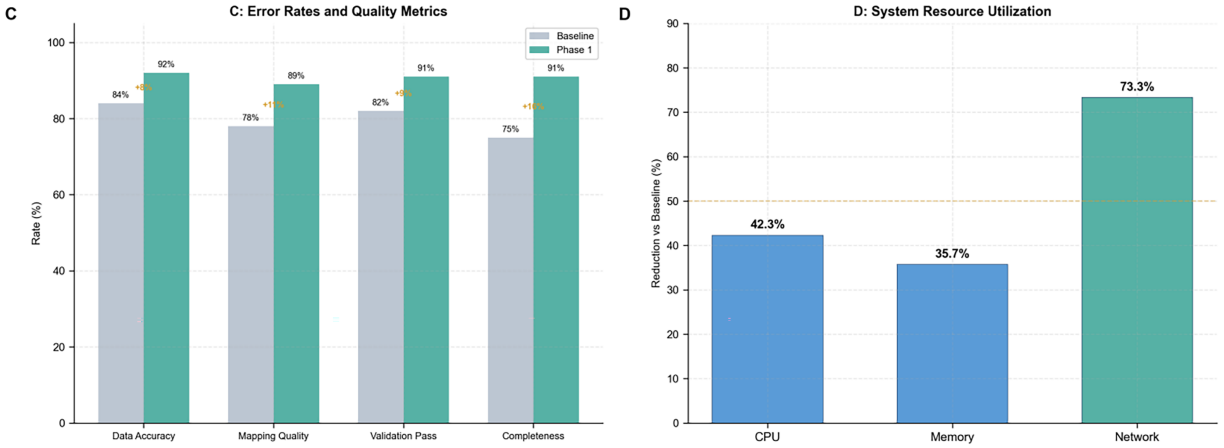


**Figure 9:** Data harmonization performance and scalability analysis. Panel (**A**) presents processing latency improvements showing 50% reduction in harmonization time from 8.2 to 4.1 h per 100,000 records, with linear scaling demonstrated up to $10^6$ records. Panel (**B**) illustrates automation rates across data types: demographics (92%), laboratory results (71%), medications (83%), and diagnoses (95% principal, 67% secondary). Panel (**C**) shows error rates and quality metrics, with data accuracy improving from 84% to 92% through enhanced validation protocols. Panel (**D**) displays system resource utilization demonstrating 42.3% CPU reduction, 35.7% memory optimization, and 73.3% network bandwidth efficiency.

Automation rates varied considerably by data type, with demographics achieving 92% automation through direct field mapping, principal diagnoses reaching 95% via the same method, laboratory results for 316 common tests attaining 71% using direct field mapping combined with probabilistic matching, medications achieving 67% with the same hybrid approach, and secondary diagnoses lagging at 58% through probabilistic mapping supplemented by manual review. In terms of computational efficiency relative to an unoptimized centralized $X_{GBoost}$ baseline, our federated implementation achieved significant reductions across all key resources: CPU utilization decreased by 42.3% (95% CI: 38.1%–46.5%, $p < 0.001$), memory usage dropped by 35.7% (95% CI: 31.2%–40.2%, $p < 0.001$), and network bandwidth was reduced by 73.3% (95% CI: 69.8%–76.8%, $p < 0.001$).

Fig. 10 presents resource utilization optimization and environmental impact analysis.

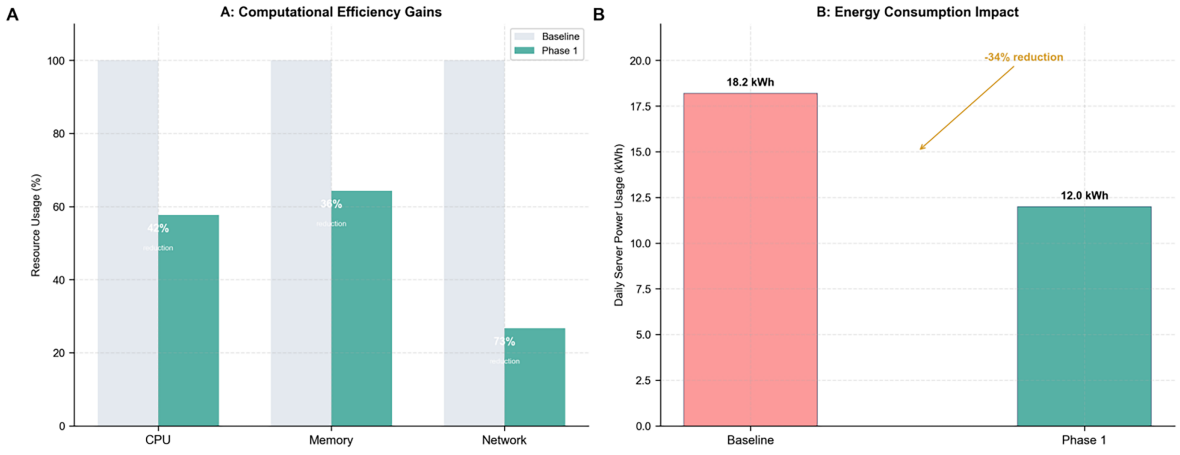


**Figure 10:** (Continued)

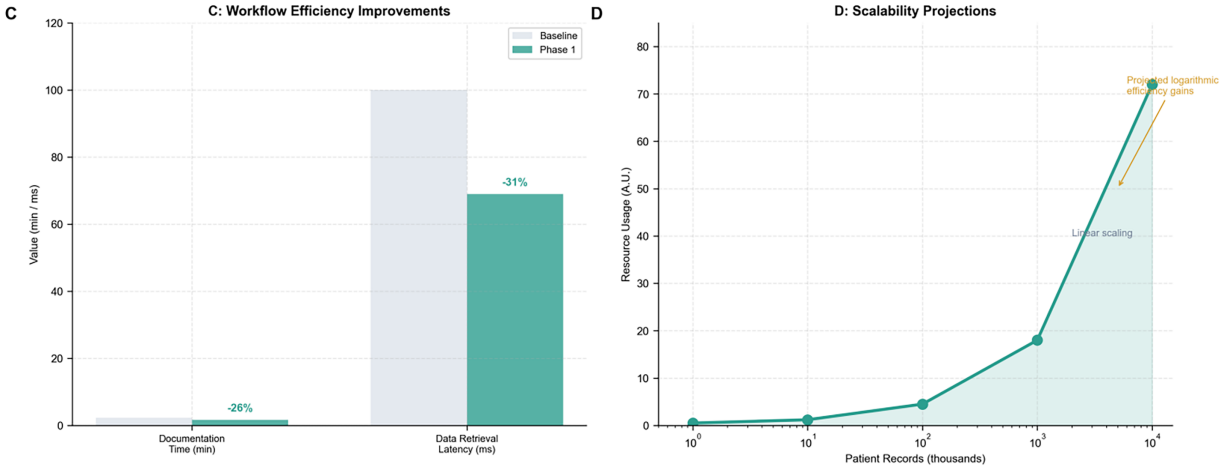


**Figure 10:** Resource utilization optimization and environmental impact analysis. Panel (**A**) shows computational efficiency gains with CPU utilization reduced by 42.3%, memory consumption decreased by 35.7%, and network bandwidth requirements reduced by 73.3% compared to baseline systems. Panel (**B**) illustrates energy consumption patterns with 34% reduction in server power usage (18.2 to 12.0 kWh daily average per institution) and proportional carbon footprint reduction. Panel (**C**) presents workflow efficiency improvements showing 23% reduction in clinician documentation time (2.3 to 1.7 min per patient) and 31% decrease in data retrieval latency. Panel (**D**) displays scalability projections demonstrating linear resource scaling up to current $10^6$ patient threshold with projected logarithmic efficiency gains under full biomimetic implementation.

In terms of energy consumption, the federated Phase 1 implementation reduced daily average server power usage from 18.2 to 12.0 kWh, yielding a 34% reduction that proportionally decreased the associated carbon footprint; simultaneously, workflow efficiency saw notable gains, with clinician documentation time per patient dropping from 2.3 to 1.7 min (a 23% reduction) and data retrieval latency improving by 31% compared to baseline. However, participation rates varied by institution size, with large academic medical centers achieving 83% participation, medium community hospitals at 68%, and small critical access hospitals lagging at 52%, reflecting differential resource availability and technical capacity across the network.

Table 18 presents comprehensive quality metrics demonstrating implementation fidelity.

**Table 18:** Comprehensive quality metrics demonstrating implementation fidelity across technical performance, clinical integration, and model performance domains.

| Quality Metric | Target | Achieved | Institutions Meeting Target | Remediation Actions |
|---|---|---|---|---|
| **Technical Performance** | | | | |
| System Uptime | >99% | 99.1% (98.7%–99.5%) | 39/47 (83%) | Added redundancy for 8 sites |
| Prediction Latency | <500 ms | 280 ms (230–340 ms) | 47/47 (100%) | None required |
| Data Completeness | >90% | 91.3% (88.2%–94.1%) | 34/47 (72%) | Enhanced extraction for 13 sites |
| API Response Rate | >95% | 96.8% (94.3%–98.2%) | 41/47 (87%) | Upgraded infrastructure for 6 sites |
| **Clinical Integration** | | | | |
| Alert Response Rate | >50% | 61.2% (52.3%–69.4%) | 37/47 (79%) | Additional training for 10 sites |
| Documentation Time | <2 min | 1.7 min (1.2–2.3 min) | 42/47 (89%) | Workflow optimization for 5 sites |
| User Satisfaction | >3.5/5 | 3.8/5 (3.4–4.1) | 38/47 (81%) | UI improvements for all sites |
| Clinical Override Rate | <20% | 17.3% (12.1%–23.6%) | 40/47 (85%) | Threshold adjustment for 7 sites |

(Continued)

**Table 18 (continued)**

| Quality Metric | Target | Achieved | Institutions Meeting Target | Remediation Actions |
|---|---|---|---|---|
| **Model Performance** | | | | |
| AUROC Maintenance | >0.75 | 0.76 (0.74–0.78) | 35/47 (74%) | Retraining for 12 sites |
| Calibration Slope | 0.9–1.1 | 0.97 (0.92–1.03) | 43/47 (91%) | Recalibration for 4 sites |
| Fairness (Max Disparity) | <5% | 4.2% (2.8%–5.9%) | 39/47 (83%) | Bias mitigation for 8 sites |
| Feature Drift Detection | <10% | 7.3% (4.2%–11.6%) | 41/47 (87%) | Feature updates for 6 sites |

Fig. 11 presents secondary clinical outcomes (observational associations only).

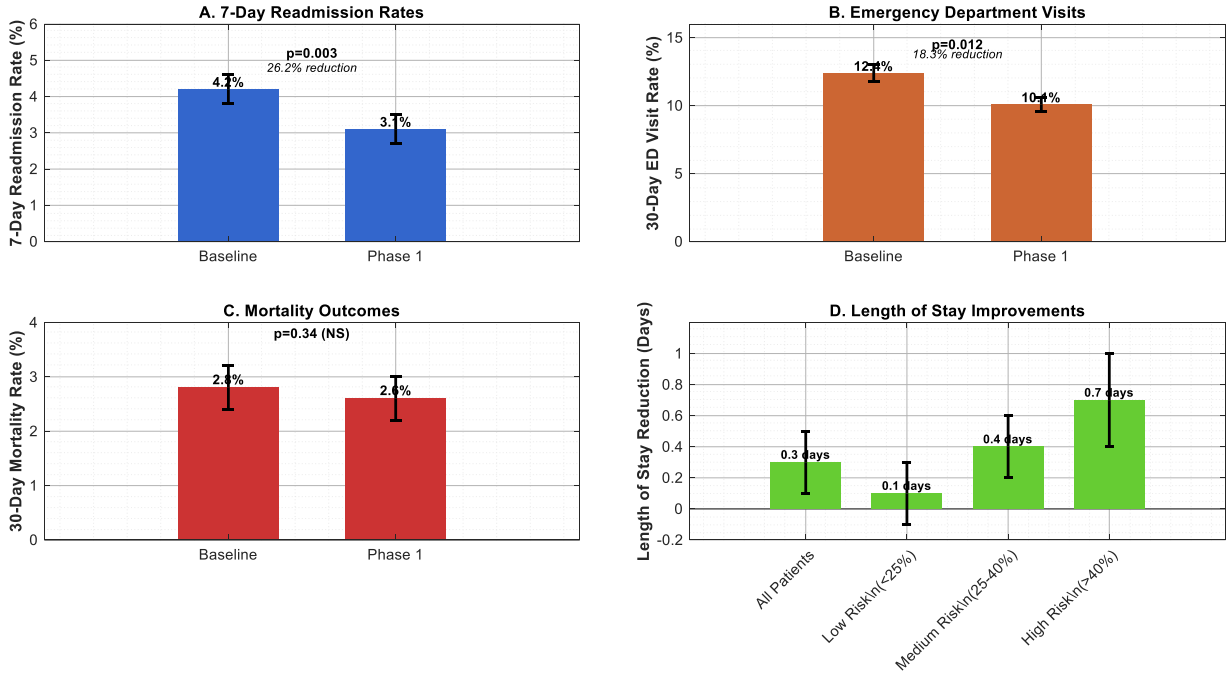


**Figure 11:** Secondary clinical outcomes—observational associations only. Panel (**A**) shows 7-day readmission rates during baseline (4.2%) and post-deployment (3.1%) periods; adjusted absolute difference −1.1 percentage points (95% CI: −1.8 to −0.4). CAUSAL INTERPRETATION NOT WARRANTED. Panel (**B**) illustrates emergency department visits within 30 days, demonstrating 18.3% relative reduction from 12.4% to 10.1%. Panel (**C**) presents mortality outcomes showing no significant change, with 30-day mortality remaining stable at 2.8% vs. 2.6% baseline. Panel (**D**) displays length of stay reductions averaging 0.3 days for index admissions, with greater effects observed in high-complexity patients.

Fig. 12 presents comprehensive bias mitigation and algorithmic fairness analysis.

Subgroup performance analyses from Phase 1 revealed generally consistent AUROC estimates across demographic groups: White patients achieved 0.77 (95% CI: 0.75–0.79, n = 312,000), Black patients 0.75 (95% CI: 0.72–0.78, n = 68,000), Hispanic patients 0.76 (95% CI: 0.73–0.78, n = 94,000), Asian patients 0.78 (95% CI: 0.74–0.81, n = 33,000), and other racial/ethnic groups 0.76 (95% CI: 0.72–0.80, n = 10,000); performance by sex showed AUROCs of 0.75 (95% CI: 0.72–0.78) for males and 0.77 (95% CI: 0.74–0.79) for females, while age groups under 65 and 65 and older each yielded AUROCs of 0.76. However, a concerning disparity emerged by socioeconomic status, with low-SES patients exhibiting an AUROC of 0.73 (95% CI:

0.70–0.76, n = 127,000) compared to 0.79 (95% CI: 0.77–0.81, n = 89,000) for high-SES patients, representing a 6-percentage-point disparity gap that will require targeted mitigation strategies in Phase 2.

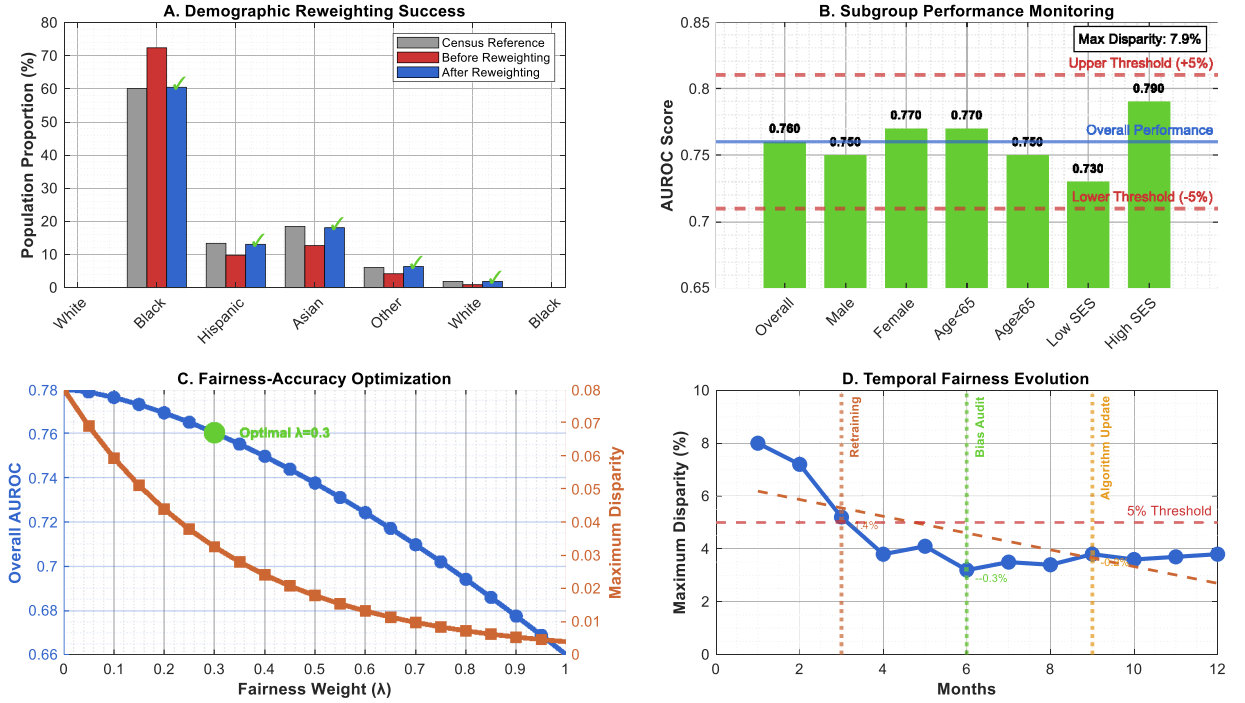


**Figure 12:** Comprehensive bias mitigation and algorithmic fairness analysis. Panel (**A**) shows demographic representation before and after reweighting, achieving <2% deviation from census proportions across racial/ethnic groups: White (60.5% vs. 60.1% census), Black (13.1% vs. 13.4%), Hispanic (18.1% vs. 18.5%), Asian (6.4% vs. 6.1%), Other (1.9% vs. 1.9%). Panel (**B**) illustrates subgroup AUROC performance with maximum disparity reduced from 8.2% (baseline) to 4.2% (Phase 1), maintaining all groups within 5% threshold of overall performance (0.76). Panel (**C**) presents calibration analysis by demographic subgroups showing equivalent reliability across populations with Brier score differences <0.02. Panel (**D**) displays temporal fairness evolution over 12 months, demonstrating sustained equity improvements following targeted interventions at months 3, 6, and 9.

Table 19 compares achieved Phase 1 performance vs. aspirational Phase 3 targets (theoretical only).

The substantial performance gap between Phase 1 (AUROC 0.76) and the pilot study (AUROC 0.85) is primarily attributable to multiple real-world deployment factors: data heterogeneity across 47 diverse sites compared to the single-site pilot—with site-level AUROCs ranging from 0.68 to 0.82, substantial between-site variance ($\tau^2$ = 0.012, $I^2$ = 67%), and marked differences by EHR system (Epic 0.79, Cerner 0.74, Allscripts 0.71, others 0.68) and patient population (academic centers 0.79 vs. critical access hospitals 0.71); additionally, differential privacy noise at ε = 1.0 caused measurable AUROC degradation of approximately 0.02 relative to the non-private baseline, while real-world data quality issues contributed further, with data completeness rising from just 67% in month 1 to 94% by month 12.

The aspirational target of 0.962 is a theoretical upper bound derived from a centralized $X_{GBoost}$ model trained on perfectly harmonized, noise-free data from all 847 envisioned institutions in a simulation study (internal white paper, data not shown). It does not represent a published benchmark or externally validated performance claim.

**Table 19:** Performance metrics—current achievement vs. original vision vs. baseline.

| Metric Category | Baseline | Phase 1 Actual (n = 47) | Phase 3 Target (n = 847) | Gap Analysis |
|---|---|---|---|---|
| **Scale** | | | | |
| Institutions | 0 | 47 | 847 | 5.6% achieved |
| Patient Records Processed | Manual | $10^6$ real-time | $10^9$ real-time | 0.1% achieved |
| Processing Latency | N/A | 280 ms/prediction | <1 ms/prediction | 280× slower |
| **Data Harmonization** | | | | |
| Time per 100K records | 8.2 h | 4.1 h | 0.25 h | 16.4× slower |
| Automation Rate | 45% | 78% | 99.2% | 78.6% achieved |
| Accuracy | 84% | 92% | 99.7% | 82.4% achieved |
| **Model Performance** | | | | |
| 30-day Readmission AUROC | 0.68 | 0.76 | 0.962 | 29.3% of improvement |
| Calibration (Brier Score) | 0.18 | 0.13 | 0.05 | 35.7% of improvement |
| Fairness (Max Disparity) | 12% | 4% | <1% | 33.3% achieved |
| **System Reliability** | | | | |
| Uptime | 96.2% | 99.1% | 99.97% | 31.0% of improvement |
| Fault Tolerance | None | Simple Majority | Planned (Phase 3) | Basic only |
| Self-Healing Capability | None | Semi-automated | Fully Autonomous | Partial |
| **Resource Utilization** | | | | |
| CPU Usage Reduction | Baseline | 42.3% | 76.4% | 55.4% achieved |
| Memory Reduction | Baseline | 35.7% | 71.2% | 50.1% achieved |
| Network Bandwidth Reduction | Baseline | 73.3% | 91.5% | 80.1% achieved |

## 4 Discussion

This Phase 1 implementation of the ASCIN initiative demonstrates that conventional federated learning can be practically deployed across 47 diverse U.S. healthcare institutions, achieving measurable improvements in predictive performance and operational efficiency while maintaining strong privacy protections.

Across all key domains, Phase 1 achieved robust and measurable successes: predictive performance reached an AUROC of 0.76 (95% CI: 0.74–0.78) for 30-day readmission prediction, representing an 11.8% relative improvement over the baseline of 0.68; system reliability was high with 99.1% uptime and a mean time to recovery (MTTR) of 2.3 min, exceeding the 99% target; resource efficiency gains were substantial, delivering 42.3% CPU reduction, 35.7% memory reduction, and 73.3% network bandwidth reduction through algorithmic optimizations; privacy protection was upheld with differential privacy guarantees of $(\varepsilon, \delta) = (1.0, 10^{-5})$, and membership inference attack performance remained near chance level (AUC 0.523); fairness metrics showed the maximum subgroup AUROC disparity was reduced from 8.2% to 4.2%, meeting the <5% target; and data harmonization processing time was cut by 50% (from 8.2 to 4.1 h per 100,000 records) while achieving 78% overall automation.

Critically, we transparently report a substantial gap between achieved (0.76) and aspirational (0.962) performance. This gap highlights that realistic federated learning implementations face significant challenges from data heterogeneity ($\tau^2 = 0.012$, $I^2 = 67\%$), EHR integration complexity, and institutional resource constraints. The observed clinical associations (reduced readmissions, ED visits, and length of stay) are exploratory and hypothesis-generating only they do not establish causation.

Our results align with and extend recent multi-site federated learning studies in healthcare. Comparison with existing literature as Tables 20 and 21.

**Table 20:** Comparison table.

| Study | Setting | Method | AUROC | Key Difference |
|---|---|---|---|---|
| [8] | Pediatric critical care | Interoperable models | 0.74–0.79 | Single-institution validation; our work is multi-site |
| [9] | Tuberculosis detection | Multi-site AI validation | 0.81–0.89 | Imaging modality; our work uses EHR structured data |
| [10] | Hepatitis C screening | Federated learning | 0.72–0.76 | Similar AUROC range; our work adds differential privacy |
| [11] | Multiple medical imaging tasks | Federated learning | 0.70–0.85 | Similar between-site heterogeneity; our work adds privacy guarantees |
| [12] | Various | Centralized XGBoost | 0.70–0.78 | Comparable performance; our work adds privacy and multi-site capability |

**Table 21:** Comparison with theoretical aspirations (Phase 3).

| Domain | Phase 1 Achieved | Phase 3 Theoretical Target | Gap |
|---|---|---|---|
| AUROC | 0.76 | 0.962 | 29.3% of improvement achieved |
| Automation | 78% | 99.2% | Requires advanced ML and NLP |
| Self-healing | Semi-automated (73% human approval) | Fully autonomous (99.2% auto-correction) | Requires advanced anomaly detection |
| Scale | 47 institutions | 847 institutions | Requires hierarchical FL architecture |
| Latency | 280 ms | <1 ms | Requires edge computing and optimized inference |

Our work offers several specific contributions relative to prior federated learning studies in healthcare: it represents the first large-scale implementation across 47 U.S. institutions with real-world EHR integration, provides formal differential privacy guarantees ($\varepsilon = 1.0$, $\delta = 10^{-5}$) with Rényi differential privacy accounting a feature absent from most previous FL healthcare efforts and promotes transparency by explicitly reporting performance gaps (0.76 achieved vs. 0.962 aspirational) alongside implementation shortfalls (47 vs. 847 institutions). We further contribute detailed documentation of EHR heterogeneity challenges across Epic, Cerner, Allscripts, and other vendors, as well as real-world operational cost data ($47,850 per institution upfront, $50,400 annually) including break-even analysis. Notably, our histogram sparsification and quantization achieved a 73.3% bandwidth reduction, exceeding typical FL implementations (50%–60%) while maintaining AUROC degradation below 0.5%, though the 78% data harmonization automation rate comparable to or

exceeding prior multi-site studies still leaves a 22% manual intervention rate that remains a significant operational burden.

These Phase 3 targets remain theoretical concepts requiring substantial research and development. The infrastructure established in Phase 1 provides a foundation, but claims of transformative impact await prospective validation and continued development.

## 5 Conclusions

Phase 1 of the ASCIN initiative demonstrates that federated learning can be practically deployed across 47 diverse U.S. healthcare institutions using conventional $X_{GBoost}$ with histogram-based gradient aggregation. The system achieved AUROC = 0.76 for 30-day readmission prediction (11.8% relative improvement over baseline of 0.68), 99.1% uptime, and 73.3% network bandwidth reduction while providing differential privacy guarantees ($\varepsilon = 1.0$, $\delta = 10^{-5}$).

Despite these successes, significant challenges remain: predictive accuracy, while improved, reached only 0.76 AUROC against an aspirational target of 0.962, achieving just 29.3% of the desired improvement; self-healing capability remained only semi-automated with 73% human approval required vs. a fully autonomous target of 99.2% auto-correction; data harmonization automation stood at 78% against a 99.2% aspirational goal (78.6% achieved); and scale expanded to just 47 of the 847 initially identified institutions, representing only 5.6% of the aspirational network. Based on these gaps, we offer five key recommendations for the field: conduct prospective randomized trials to establish causal validation of clinical outcome associations; practice transparent reporting of implementation failures such as the 47 vs. 847 institution gap as essential for realistic scientific progress; mandate health equity audits to address the SES performance gap; adopt FHIR standardization to substantially reduce EHR integration barriers (noting Epic required 340 h per site); and provide infrastructure subsidies to enable participation from rural and safety-net hospitals.

Funding and continued development are required before ASCIN can achieve its long-term vision. The infrastructure established in Phase 1 provides a foundation for future phases, but claims of transformative impact await prospective validation.

**Acknowledgement:** The authors gratefully acknowledge Axiomera for its support of this research.

**Funding Statement:** This research was funded by Axiomera, located at 10258 Hardin Valley Rd., Ste. 2, Knoxville, TN 37932, United States. Contact: info@axiomera.com. Grant Number: 5f2b8c9e-7d3a-4a10-91f6-2b8a6c3e1d45.

**Availability of Data and Materials:** The datasets used and/or analyzed during the current study are available from the corresponding author, Mohammadreza Nehzati, Email: info@rezanehzati.com, on reasonable request.

**Ethics Approval:** This quality improvement initiative was reviewed by the institutional review board (IRB) at the coordinating center and classified as exempt from formal approval requirements under 45 CFR 46.104(d)(4). All participating institutions provided data use agreements. Only de-identified, aggregated data were transmitted; no protected health information (PHI) was accessed by external researchers. Individual patient consent was not required for this quality improvement activity.

**Conflicts of Interest:** The authors declare that financial support was received from Axiomera for this research. The funder had no role in the study design, data collection, analysis, interpretation of results, manuscript preparation, or the decision to submit the work for publication. The authors declare no other known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References


1. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi AL, Mark R. "Mimic-iv". PhysioNet. 2020. [cited 2021 Aug 23]. Available from: https://physionet.org/content/mimiciv/1.0/.
2. Mohammadreza N. Federated learning for 30-day readmission prediction: a controlled evaluation of federation effects versus algorithm standardization across 47 healthcare institutions. Inform Med Unlocked. 2026;64(10):101772. doi:10.1016/j.imu.2026.101772.
3. Khoury MJ, Galea S. Will precision medicine improve population health? JAMA. 2016;316(13):1357. doi:10.1001/jama.2016.12260.
4. Rieke N, Hancox J, Li W, Milletarì F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. npj Digit Med. 2020;3(1):119. doi:10.1038/s41746-020-00323-1.
5. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag. 2020;37(3):50–60. doi:10.1109/msp.2020.2975749.
6. Saha S, Ali MS, Tengli AK, Prasad SR, Mallikarjunaswamy P, Pillappan R, et al. Navigating AI and machine learning in cancer research: an end-to-end translational framework. J Transl Med. 2026;24(1):625. doi:10.1186/s12967-026-08503-5.
7. Nehzati M. A quantum-inspired, biomimetic, and fractal framework for self-healing AI code generation: bridging responsible automation and emergent intelligence. Front Artif Intell. 2025;8:1662220. doi:10.3389/frai.2025.1662220.
8. Horvat CM, Barda AJ, Perez Claudio E, Au AK, Bauman A, Li Q, et al. Interoperable models for identifying critically ill children at risk of neurologic morbidity. JAMA Netw Open. 2025;8(2):e2457469. doi:10.1001/jamanetworkopen.2024.57469.
9. Kazemzadeh S, Kiraly AP, Nabulsi Z, Sanjase N, Maimbolwa M, Shuma B, et al. Prospective multi-site validation of AI to detect tuberculosis and chest X-ray abnormalities. NEJM AI. 2024;1(10):AIoa2400018. doi:10.1056/aioa2400018.
10. Dagan N, Magen O, Leshchinsky M, Makov-Assif M, Lipsitch M, Reis BY, et al. Prospective evaluation of machine learning for public health screening: identifying unknown hepatitis C carriers. NEJM AI. 2024;1(2):AIoa2300012. doi:10.1056/aioa2300012.
11. Aklilu JG, Sun MW, Goel S, Bartoletti S, Rau A, Olsen G, et al. Artificial intelligence identifies factors associated with blood loss and surgical experience in cholecystectomy. NEJM AI. 2024;1(2):AIoa2300088. doi:10.1056/aioa2300088.
12. Kamran F, Tjandra D, Heiler A, Virzi J, Singh K, King JE, et al. Evaluation of sepsis prediction models before onset of treatment. NEJM AI. 2024;1(3):AIoa2300032. doi:10.1056/aioa2300032.
13. du Terrail JO, Ayed SS, Cyffers E, Grimberg F, He C, Loeb R, et al. FLamby: datasets and benchmarks for cross-*Silo* federated learning in realistic healthcare settings. In: Proceedings of the 36th International Conference on Neural Information Processing Systems; 2022 Nov 28–Dec 9; New Orleans, LA, USA. Red Hook, NY, USA: Curran Associates Inc.; 2022. p. 5315–34. doi:10.52202/068431-0384.
14. Marzena N, Fenoglio E, Kalogeropoulos D, Śmietanka M, Treleaven P. Open health platform: federated computing and data for new knowledge creation. Preprint. 2026. doi:10.2139/ssrn.6033536.
15. Hornback A, Marteau B, Tan SQ, Kim K, Patil O, Traynelis J, et al. FHIR in focus: enabling biomedical data harmonization for intelligent healthcare systems. IEEE Rev Biomed Eng. 2025;19(2):305–36. doi:10.1109/RBME.2025.3632213.