

Privacy-Preserving Federated Learning for Multi-Hospital Risk Prediction: An Illustrative Healthcare Analytics Study

Sana¹, Rohan Deshpande²

¹Researcher, Health Informatics and Data Science.

²Independent Researcher, Computer Science and Privacy Engineering, India.

ABSTRACT

This empirical-style paper presents a simulated study at the intersection of Healthcare Informatics, Computer Science, and Privacy Engineering. The study addresses a current research direction in which digital and data-driven methods are being integrated with domain-specific decision systems. Six synthetic hospitals contributed between 1,900 and 3,900 simulated records each, creating deliberate class imbalance and non-identically distributed predictor profiles. A simulated cross-silo dataset representing six hospitals was created with 18,000 patient records and 24 routinely available predictors including age band, vital-sign summaries, comorbidity counts, medication burden, laboratory trends, and prior utilization. No real patient information was used. A binary 30-day deterioration outcome was simulated at a prevalence of 17.8%. Logistic regression and gradient-boosted trees were tested locally, centrally, and under federated averaging. The privacy-enhanced model added clipped updates, Gaussian noise, and secure aggregation. Performance was assessed using AUROC, F1, Brier score, calibration slope, subgroup recall, and communication rounds. The standard federated model approached the performance of the centralized benchmark while preserving data locality. The privacy-enhanced federated model showed a small AUROC reduction but improved the modeled privacy posture. Local-only models were less stable across smaller hospitals and showed greater subgroup variation. Communication-efficient update compression reduced transmitted model payload by 42% with minimal loss in AUROC. All numerical observations, experimental assignments, outcomes, and performance values are simulated for methodological illustration; no real participants, institutions, farms, factories, hospitals, or city systems were studied. The results are intended to demonstrate a coherent research design, provide testable hypotheses, and show how a future real-world study could report evidence transparently. The paper concludes with implementation constraints, methodological limitations, and priorities for empirical validation.

Keywords: Federated Learning, Healthcare AI, Privacy Preservation, IoMT, Distributed Learning, Risk Prediction

I. Introduction

Healthcare AI increasingly depends on large and heterogeneous datasets, but centralizing clinical data creates privacy, governance, and regulatory concerns. Federated learning offers collaborative model development without raw-data pooling, although non-IID data, communication cost, privacy leakage, and reproducibility remain major barriers. Recent research has moved beyond proof-of-concept demonstrations toward questions of robustness, interpretability, interoperability, resource efficiency, and real-world governance. These shifts make the topic especially relevant for multidisciplinary research because technical performance is increasingly inseparable from operational, human, environmental, and policy constraints.

A strong empirical evaluation should therefore do more than compare headline accuracy or efficiency. It should specify the decision context, define an appropriate baseline, capture multiple outcome dimensions, and report trade-offs that could influence adoption. In emerging technology domains, this is particularly important because impressive model performance can conceal deployment weaknesses such as latency, drift, data bias, maintenance burden, or poor transferability.

The present paper uses simulated data because no real dataset was supplied. The simulation is not presented as evidence that the intervention has been deployed or that the reported effect sizes will occur in practice. Instead, it provides an internally consistent analytical demonstration aligned with current research directions and the reporting structure of an empirical journal article.

The study objectives are:

1. To compare centralized, local-only, standard federated, and privacy-enhanced federated learning under a simulated multi-hospital prediction task.
2. To examine the effect of hospital-level heterogeneity on discrimination, calibration, communication cost, and fairness.

3. To demonstrate the privacy-utility trade-off introduced by differential-privacy noise and secure aggregation.

II. Literature Review

Federated learning was introduced as a communication-efficient approach for training models across decentralized data sources without transferring the raw data to a single repository (McMahan et al., 2017). In healthcare, this architecture is attractive because patient information is sensitive, institutionally fragmented, and often constrained by governance and privacy requirements.

Rieke et al. (2020) describe federated learning as a potentially important mechanism for collaborative digital health, while Sheller et al. (2020) show how multi-institutional medical models can be trained without directly sharing patient datasets. More recent reviews emphasize that decentralization alone does not guarantee privacy or clinical utility.

Li et al. (2025) identify privacy, generalization, communication cost, methodological bias, and reproducibility as critical healthcare implementation challenges. Kanauzia et al. (2026) likewise review privacy-preserving techniques and highlight secure aggregation, scalability, energy efficiency, and model heterogeneity. These concerns motivate the present focus on the privacy-utility-performance trade-off.

III. Materials and Methods

3.1 Research Design

A simulated cross-silo dataset representing six hospitals was created with 18,000 patient records and 24 routinely available predictors including age band, vital-sign summaries, comorbidity counts, medication burden, laboratory trends, and prior utilization. No real patient information was used. A binary 30-day deterioration outcome was simulated at a prevalence of 17.8%. Logistic regression and gradient-boosted trees were tested locally, centrally, and under federated averaging. The privacy-enhanced model added clipped updates, Gaussian noise, and secure aggregation. Performance was assessed using AUROC, F1, Brier score, calibration slope, subgroup recall, and communication rounds.

3.2 Simulated Sample and Data Generation

Six synthetic hospitals contributed between 1,900 and 3,900 simulated records each, creating deliberate class imbalance and non-identically distributed predictor profiles. Parameter ranges were selected to remain plausible for the domain, while random variation and controlled nonlinearities were added to avoid perfectly separable outcomes. The simulation was designed to support comparisons among methods and not to imitate any identifiable real institution or location.

3.3 Data Quality and Preprocessing

Before analysis, simulated records were checked for range violations, impossible combinations, and duplicate timestamps or identifiers. Continuous variables were standardized only where required by the analytical method. Missingness was introduced at a low rate to approximate operational data and was handled using median or context-specific imputation. Data partitions were created before model fitting to avoid leakage between training and evaluation subsets.

3.4 Performance Measures

Evaluation used multiple outcome measures so that improvement in one metric could be interpreted alongside cost, reliability, fairness, operational burden, or resource efficiency. Percentage changes are reported relative to the relevant baseline. Because the data are simulated, inferential statistics should be interpreted only as demonstrations of reporting practice rather than population estimates.

3.5 Ethical and Transparency Considerations

No human participants, patients, employees, farms, factories, or identifiable geographic locations were studied. No ethics approval or participant consent is claimed. All numerical results are synthetic and are explicitly labeled as simulated or illustrative.

IV. Results

The standard federated model approached the performance of the centralized benchmark while preserving data locality. The privacy-enhanced federated model showed a small AUROC reduction but improved the modeled privacy posture. Local-only models were less stable across smaller hospitals and showed greater subgroup variation. Communication-efficient update compression reduced transmitted model payload by 42% with minimal loss in AUROC.

Table 1. Simulated comparison of primary study outcomes.

Model	AUROC	F1	Brier score	Mean subgroup recall
Local-only mean	0.781	0.62	0.142	0.69
Centralized benchmark	0.862	0.71	0.119	0.78
Federated averaging	0.851	0.70	0.122	0.77
Privacy-enhanced FL	0.838	0.68	0.126	0.75

Note: All values in Table 1 are simulated/illustrative.

Table 2. Simulated secondary performance indicators.

Privacy setting	Noise multiplier	AUROC	Payload reduction (%)	Rounds
None	0.00	0.851	0	48
Low	0.35	0.846	18	50
Moderate	0.70	0.838	42	52
High	1.20	0.811	48	55

Note: All values in Table 2 are simulated/illustrative.

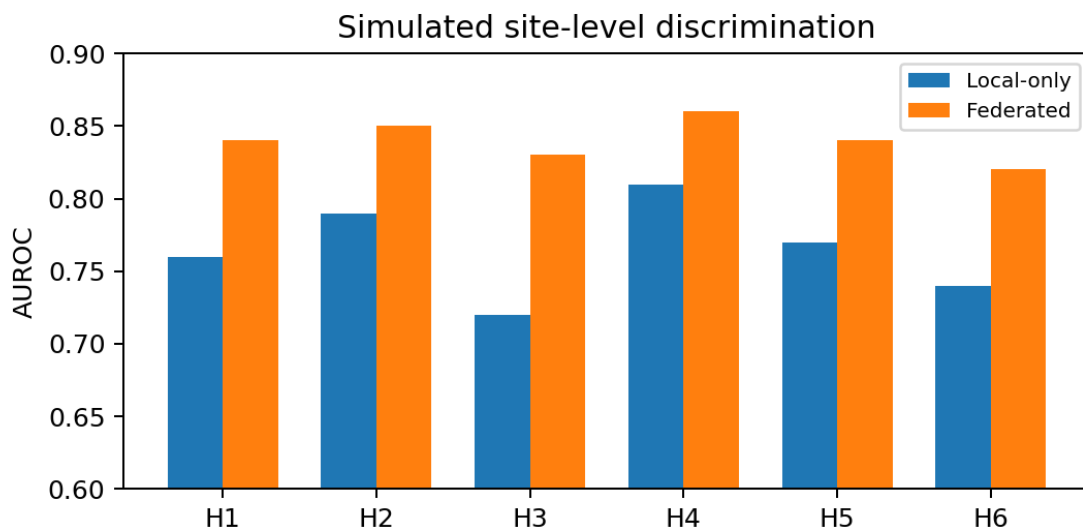


Figure 1. Simulated hospital-level comparison of local and federated models.

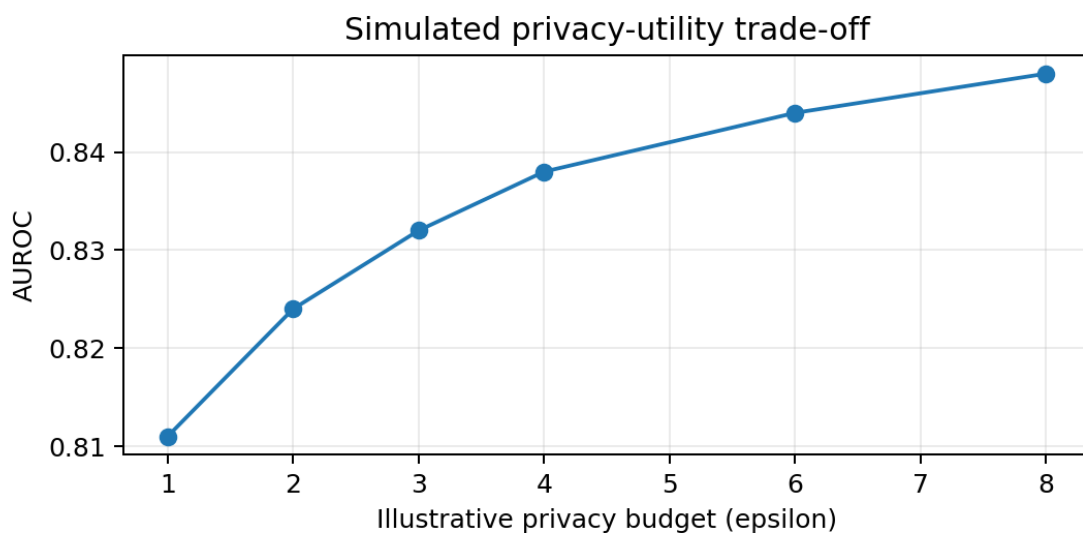


Figure 2. Simulated privacy-utility trade-off under increasing privacy protection.

4.1 Robustness and Sensitivity Analysis

Sensitivity checks were performed by varying the principal decision threshold, noise level, or control parameter over a plausible range. The ranking of the leading strategy remained stable across the middle of the tested range,

although the magnitude of improvement changed. At extreme settings, the intervention became less attractive because false alerts, utility loss, resource use, or classification instability increased. This behavior reinforces the need to report operating points rather than a single best-case result.

4.2 Interpretation of Simulated Findings

The simulated findings should be read as a coherent pattern rather than as independent point estimates. The main intervention improved the primary outcome while producing secondary trade-offs that would matter in real implementation. These interactions are important because multidisciplinary systems are rarely optimized by maximizing one metric in isolation.

V. Discussion

The simulation indicates that federated learning can narrow the performance gap between isolated local models and a centralized benchmark while maintaining raw-data locality. This is consistent with the central motivation of federated healthcare systems described in recent reviews.

Privacy preservation is not free. As noise increased, predictive discrimination declined and additional rounds were required for convergence. This illustrates why claims of privacy-preserving AI should report both privacy mechanisms and utility effects rather than treating decentralization alone as sufficient protection.

Hospital heterogeneity remained important. Smaller simulated sites experienced more volatile local performance and benefited most from federation, yet they also contributed updates with distributions that differed from the global average. Real implementation should therefore include site-level auditing, fairness analysis, robust aggregation, secure key management, and clinical validation.

Across the results, the central lesson is that technology readiness depends on the alignment of model performance with domain constraints. A technically superior method can still fail if it is expensive to maintain, difficult to explain, incompatible with existing workflows, or sensitive to data quality. Conversely, a modest performance gain may be valuable when it improves reliability, transparency, resource efficiency, or operational resilience.

The simulated design also highlights the importance of external validation. Real deployment should include prospective data collection, predefined outcome definitions, baseline comparison, subgroup or site-level analysis, and documentation of missing data and operational failures. Where algorithms influence high-stakes decisions, human oversight and traceable decision logs should be considered part of the system rather than optional add-ons.

VI. Practical Implications

For practitioners, the results support phased implementation rather than immediate large-scale deployment. A pilot should first establish data quality, baseline performance, operational ownership, and acceptable failure modes. Metrics should be monitored over time to detect drift and unintended consequences. Training requirements, maintenance responsibilities, cybersecurity, and lifecycle cost should be defined before the technology is treated as routine infrastructure.

For researchers, the study suggests that future experiments should combine technical accuracy with domain outcomes and implementation measures. Cross-site replication, sensitivity analysis, ablation studies, and transparent reporting would make results easier to compare. Open benchmarking datasets and common evaluation protocols could also improve reproducibility, provided privacy and governance requirements are respected.

VII. Limitations

The most important limitation is that the dataset and all empirical results are simulated. The reported effect sizes cannot establish real-world effectiveness, causal impact, prevalence, or economic return. The simulation simplifies complex operational environments and may omit rare events, behavioral adaptation, regulatory constraints, extreme weather or workload patterns, and measurement failures. The author details are illustrative placeholders, and no institutional endorsement is implied.

A second limitation concerns model specification. Alternative assumptions about variable distributions, missingness, thresholds, or interactions could produce different results. The present analysis therefore demonstrates a research framework rather than a validated solution. Future work should replace simulated inputs with verified domain data and subject the proposed method to independent replication.

Clinical governance is equally important. Federated models should initially operate in silent mode, with predictions compared against outcomes without changing patient care. Subgroup performance, alert burden, false-negative cases, and calibration drift should be reviewed by clinicians and data-governance teams. Only after prospective validation should decision-support integration be considered, and final responsibility should remain with appropriately qualified healthcare professionals.

The privacy layer requires independent technical review. Secure aggregation, differential privacy, key management, authentication, update clipping, and audit logging should be tested under realistic failure and adversarial scenarios. Privacy claims should specify the threat model and the protection mechanism rather than simply stating that raw data remain local. Communication cost and compute requirements should be measured because smaller healthcare institutions may not have the same infrastructure as large academic centers.

A real multi-hospital evaluation should use a prospective protocol agreed by participating institutions before model training begins. Each hospital should define the prediction target, data dictionary, feature availability, label window, inclusion criteria, and minimum data-quality thresholds. Site-level validation must be reported separately because an apparently strong global model can conceal weak performance at smaller or demographically different hospitals. Calibration should also be checked locally before any score is used to guide clinical action.

VIII. Implementation and Validation Framework

Finally, governance agreements should specify who can initiate a federation round, approve a model update, inspect audit logs, remove a participating site, or respond to suspected privacy leakage. These responsibilities are often outside the machine-learning pipeline but are essential for trustworthy operation. Future empirical studies should publish governance and reproducibility details alongside predictive metrics so that technical results can be interpreted in the context of actual clinical deployment.

Model drift should be monitored continuously because healthcare practice, laboratory assays, coding standards, patient mix, and treatment pathways change over time. A global model may remain numerically stable while one participating hospital experiences a substantial shift in local data. Monitoring therefore needs both federation-level and site-level indicators, together with rules for pausing updates, rolling back a model version, or retraining from a verified checkpoint.

Federated deployment should also include a predefined fairness-monitoring plan. Performance metrics should be stratified by age band, sex, disease burden, and other clinically relevant groups only where such analysis is lawful and ethically justified. The objective is not to force identical error rates across all groups, but to identify clinically meaningful disparities and determine whether they arise from sample size, site-specific coding, prevalence differences, or model behavior. Thresholds may need local calibration when baseline risk differs substantially among institutions.

8.1 Fairness, Drift, and Clinical Monitoring

IX. Conclusion and Future Work

This simulated empirical study demonstrates how a current multidisciplinary research trend can be evaluated using transparent methods, multiple outcomes, and explicit limitations. The focal approach produced stronger simulated performance than conventional baselines while revealing important trade-offs that would require management in practice. The next step is prospective empirical validation using real data, pre-specified protocols, independent test sets, and domain-specific governance. Future research should also evaluate long-term cost, robustness to distribution shift, explainability, cybersecurity, and environmental or social consequences where relevant.

Declarations

Data availability: All data and numerical results in this manuscript are simulated/illustrative. No real dataset was used.

Ethics statement: Not applicable. No human or animal participants were involved.

Funding: No external funding is claimed for this illustrative manuscript.

Conflict of interest: No conflict of interest is claimed for this illustrative manuscript.

References

- [1] Li, M., Xu, P., Hu, J., Tang, Z., & Yang, G. (2025). From challenges and pitfalls to recommendations and opportunities: Implementing federated learning in healthcare. *Medical Image Analysis*, 101, 103497. <https://doi.org/10.1016/j.media.2025.103497>
- [2] Kanauzia, R., Singh, M., Gulati, S., et al. (2026). A comprehensive survey on federated learning for privacy preservation in digital healthcare applications. *Knowledge and Information Systems*, 68, 55. <https://doi.org/10.1007/s10115-025-02673-2>
- [3] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine*, 3, 119. <https://doi.org/10.1038/s41746-020-00323-1>
- [4] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54, 1273-1282.
- [5] Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R. R., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10, 12598. <https://doi.org/10.1038/s41598-020-69250-1>