Federated Learning for Privacy-Preserving Healthcare Data Analysis

Tanya Soni, M. Tech Scholar, Department of Computer Science & Engineering, Swami Vivekanand Institute of Engineering and Technology, Banur, Punjab, India, sonitanya0019@gmail.com

Prof. (Dr.) Arun Kumar, Professor, Department of Computer Science & Engineering, Swami Vivekanand Institute of Engineering and Technology, Banur, Punjab, India, arunkumarsai@gmail.com

Abstract — Hospitals hold large amounts of patient data, but privacy laws such as HIPAA, GDPR and the Indian DPDP Act make it very difficult to pool this data for research. Federated learning offers a way out: each hospital trains on its own records and shares only model parameters. This paper measures what that arrangement actually costs. Using two real clinical datasets, FLCHAIN (7,874 Mayo Clinic subjects) and the Breast Cancer Wisconsin dataset (569 cases), we ran 768 seeded experiments across ten simulated hospitals under a single, fully stated privacy accounting convention. Federated averaging reached an AUROC of 0.8319 against 0.8389 for centrally pooled data, retaining 99.2 per cent of the accuracy, while hospitals working alone reached only 0.7851. Adding differential privacy proved far more expensive than data heterogeneity: 96.4 per cent of accuracy survived at a privacy budget of 10, but only 79.3 per cent at 1. Three practical findings are reported. Adding a proximal term to the local objective halves the privacy budget needed for the same accuracy. A membership inference attack scored only 0.510 against a model with no privacy protection, so the formal guarantee is very conservative. Finally, privacy noise harms the weakest patient subgroup about 2.1 times faster than average accuracy.

Keywords — ***Clinical risk prediction, Differential privacy, Electronic health records, Fairness, Federated learning, Membership inference.***

I. INTRODUCTION

Machine learning is now used widely for clinical decision support, but the data it needs is locked inside individual hospitals. Patient records are collected under consent agreements written on the assumption that the data will stay where it was collected, and laws such as HIPAA in the United States, GDPR in Europe and the Digital Personal Data Protection Act of 2023 in India restrict transfers of identifiable health information between organisations. Surveys of the field confirm that this fragmentation is the main practical barrier to building clinical models that generalise well [1], [2].

The cost of working alone is measurable. Dayan et al. trained a model across twenty institutions and reported a 38 per cent average improvement in generalisability compared with single-site models [3]. Sheller et al. found the same pattern across ten institutions and concluded that the extra data gained through collaboration helps more than the error introduced by the collaborative method itself [4].

Federated learning solves the access problem by moving the model to the data instead of the data to the model [5]. A central server sends model parameters to each hospital, each hospital trains on its own records, and the server averages the returned parameters. No record crosses an institutional boundary. Fig. 1 shows the arrangement used in this study.

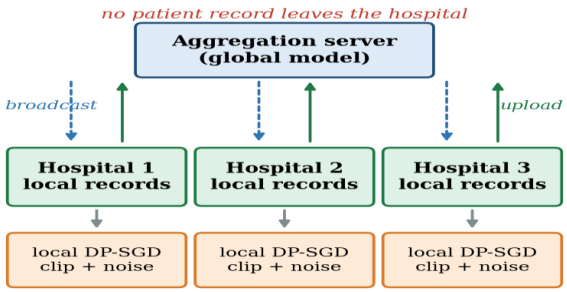

Fig. 1. Architecture of the federated learning system evaluated in this study. The aggregation server holds the global model and exchanges only parameters with the participating hospitals; each hospital applies local DP-SGD, clipping every per-example gradient and adding Gaussian noise before its update is uploaded. No patient record leaves the hospital at any stage.

 DOI : 10.35291/2454-9150.2026.0044

Fig. 1 identifies the three elements that the remainder of this paper measures. The downward arrows carry the broadcast parameters, the upward arrows carry the local parameter updates, and the shaded boxes between them mark the point at which differential privacy is applied. Placing the noise at the client rather than at the server is the design choice that makes the guarantee independent of any trust in the aggregator, and it is also the choice that makes privacy expensive, because the noise of every client enters the average. The outcome of that choice is quantified in Section IV-C.

However, federated learning is not private by itself. Model updates are functions of the training data and are known to leak information. Shokri et al. demonstrated membership inference against models trained on hospital discharge data [6], and Hu et al. showed that a curious server can even identify which hospital contributed a record, with 67.1 per cent success against a 10 per cent baseline [7]. Differential privacy is the standard formal remedy [8], but it reduces accuracy, and in medicine that loss must be measured before it can be accepted.

Measuring it from published work is difficult because reported privacy budgets are not comparable. Studies quote an accuracy at a stated value of epsilon without saying whether the guarantee protects one patient record or an entire hospital, whether it is a worst case or an average, or what value of delta was used. Reported operating points therefore range widely, from 1.9 [9] to about 9.0 [10], and a recent review of 74 studies notes that reporting of privacy parameters remains inconsistent [11].

A. Contributions

This paper measures the cost of privacy in federated clinical prediction on real patient data, under one clearly stated convention, with variation reported across repeated runs. The main contributions are:

1) A separation of the value of collaboration from the cost of federation on a cohort of 7,874 patients, showing that the gain from collaborating is about seven times larger than the loss from federating.

2) The finding that adding a proximal term to the local objective halves the privacy budget needed for a given level of accuracy.

3) An attack-based audit showing that a membership inference adversary gains far less than the formal guarantee allows.

4) Evidence that differential privacy harms the weakest patient subgroup about twice as fast as it harms average accuracy.

5) A fully specified experimental protocol, including the alternatives rejected at each design point, so that the measurements can be reproduced and contested.

II. RELATED WORK

Federated averaging remains the reference algorithm [5]. Clients run local stochastic gradient descent and the server takes a weighted mean of the returned parameters. Li et al. provided the convergence analysis for data that is not identically distributed and showed that heterogeneity slows convergence [12]. FedProx adds a proximal penalty that discourages the local model from drifting away from the broadcast parameters, and reports accuracy gains averaging 22 per cent in highly heterogeneous settings [13].

On the privacy side, Wei et al. gave the foundational analysis of noisy federated aggregation and proved an explicit trade-off between convergence quality and protection level [8]. Ziller et al. released an open framework for differentially private training and demonstrated it on paediatric pneumonia and liver segmentation tasks [14]. Accounting for the Poisson sub-sampled Gaussian mechanism follows Wang et al. [15]. Bai et al. survey the attacks and defences specific to the federated setting [16].

Several refinements of the basic scheme target specific kinds of heterogeneity. FedBN keeps normalisation statistics on the client instead of averaging them, which suits the situation where different hospitals use different scanners or assays [21]. Manthe et al. benchmarked global, personalised and hybrid schemes on a brain tumour segmentation task and found that plain federated averaging is already close to optimal, with the more elaborate methods adding only small gains [22]. This influenced our decision to use federated averaging and FedProx as the working baselines rather than a more complex scheme.

Cryptographic methods are often presented as alternatives to differential privacy, but they act at a different point in the pipeline. Secure aggregation and homomorphic encryption stop the server from reading any individual client update, yet they say nothing about what the final released model reveals, because that model is the intended output of the protocol. Differential privacy does the opposite. Firdaus et al. combine blockchain with homomorphic encryption to reduce the trust placed in the central server [23], and the MetisFL system of Stripelis et al. computes the global model under full encryption while also limiting information leakage from the model itself [24]. Since encryption does not change the learning dynamics, it cannot alter accuracy, and it is therefore discussed here rather than tested experimentally.

Fairness under privacy has received much less attention. The review by Mohammadi et al. observes that only a minority of studies evaluate fairness at all, and that several of those that do report differential privacy widening performance gaps between patient subgroups [11]. Section IV-E addresses this gap directly.

The most recent literature continues along these same three lines and sharpens the questions asked here. On the clinical side, Hasan et al. combine electronic health records with ECG

 DOI : 10.35291/2454-9150.2026.0044

time series in a multi-modal federated model trained under differential privacy, reporting 94.1 per cent accuracy and convergence 32.4 per cent faster than single-modality training while remaining stable under heterogeneous client distributions [25], and Lee et al. describe an international federated platform for paediatric brain tumours that demonstrates the arrangement is workable across real hospital governance boundaries rather than only in simulation [20]. On fairness, the FlexFair framework of Xing et al. introduces a flexible regularisation term that allows equal accuracy, demographic parity and equal opportunity to be optimised together across four clinical imaging tasks [26]; this is the closest existing response to the subgroup disparity measured in Section IV-E, although it is not evaluated under a differential privacy budget, so the interaction between the two mechanisms remains open. On the audit side, Annamalai and De Cristofaro obtain a nearly tight black-box audit of DP-SGD and recover an empirical epsilon of about 7.0 where the theoretical budget is 10 [27]. The gap they measure on image data is the same phenomenon reported in Section IV-F, and their result supports the argument made there that the formal budget is a conservative upper bound rather than a description of the leakage a hospital actually faces.

III. MATERIALS AND METHODS

A. Datasets

The main cohort is FLCHAIN, drawn from the Mayo Clinic study of serum free light chains reported by Dispenzieri et al. [17]. The parent study enrolled residents of Olmsted County, Minnesota aged 50 and above, with samples collected between 1995 and 2003 and survival followed to 2009. The redistributable subset used here contains 7,874 subjects with complete assay data. The task is binary prediction of all-cause mortality, with an event rate of 27.55 per cent.

Two variables were removed before modelling. The recorded cause-of-death code appears only for subjects who died and would therefore leak the label completely. The enrolment year was also withheld from the feature set, because it identifies the client. Creatinine, missing for 1,350 subjects, was replaced by the training median with an added missingness flag, and four clinically meaningful features were derived: the logarithms of kappa and lambda, their sum and their ratio. This gives thirteen predictors, standardised using training-split statistics only.

The second cohort is the Breast Cancer Wisconsin diagnostic dataset of 569 cases with thirty features [18], used because published federated results with differential privacy already exist for it.

B. Building the federation

Hospitals were simulated using Dirichlet partitioning, in which a proportion vector drawn from a Dirichlet distribution with concentration alpha divides the records of each class among K clients. Rather than describing a partition only by alpha, we also report the mean total variation distance between each client's outcome rate and the pooled rate:

$$TV = (1/K) \Sigma \mid p_k(y = 1) - p(y = 1) \mid \qquad (1)$$

This number can be computed by a real hospital group from summary statistics alone, without sharing any records, so a consortium can locate itself on the curves reported here before joining a project.

Synthetic partitions are convenient but artificial, so we also built a federation that is not synthetic. FLCHAIN records the year in which each subject was enrolled, and splitting on that variable gives nine sites containing 949, 2,594, 1,057, 527, 257, 189, 131, 35 and 166 subjects for the years 1995 to 2003. These groups differ in size by a factor of 74 and in mortality rate from 2.9 to 31.7 per cent, yet their total variation distance is only 0.085. This is a useful calibration point: real institutional splits appear to differ mainly in how much data each site holds, rather than in the mix of outcomes, and are therefore milder than the synthetic partitions commonly used in this literature.

C. Model and algorithms

All experiments use a small neural network with one hidden layer of 32 rectified linear units and a sigmoid output, giving 481 parameters. The network was chosen deliberately. Per-example gradients are available in closed form for this shape, which makes rigorous private training possible without an automatic differentiation library, and a small parameter count keeps the noise scale manageable, since noise is added to every coordinate.

Four strategies were compared with two reference points. Federated averaging takes the weighted mean of client parameters. FedProx adds a proximal penalty with coefficient mu. FedAvgM applies server momentum. The centralised baseline trains on pooled data and gives the upper bound if privacy were not a constraint, while the isolated baseline trains each hospital separately. Unless stated otherwise, the setting is ten clients, sixty rounds, two local epochs, batch size 32 and learning rate 0.15, with five independent seeds.

D. Differential privacy

Each client runs differentially private stochastic gradient descent locally. A Poisson sample is drawn at each step, the per-example gradient of every sampled record is clipped to a fixed L2 norm C, the clipped gradients are summed, Gaussian noise of standard deviation sigma times C is added, and the result is divided by the expected batch size. Because clipping bounds the influence of any single record before noise is applied, the sensitivity of the sum is exactly C.

Privacy is accounted using the Rényi accountant for the Poisson sub-sampled Gaussian mechanism [15], composed over every local step across all rounds, at a delta of ten to the power minus five. Two choices are stated explicitly because most papers leave them implicit. The guarantee is at the level of a single patient record, not an entire hospital. The reported

 DOI : 10.35291/2454-9150.2026.0044

budget is the worst case across clients, not the average, so it holds for every patient including those at the smallest site.

E. Evaluation

AUROC is the primary measure, with AUPRC reported alongside because the outcome is imbalanced. Realised leakage is measured with a loss-threshold membership inference attack, in which the adversary scores each candidate record by the negative of its per-example loss; a score of 0.5 means no leakage. Fairness is assessed by splitting the test set by sex and into three age bands. In total 768 training runs were carried out and every figure is the mean of five seeds.

F. Experimental protocol and tools

The complete pipeline runs in seven steps, executed identically for every configuration and every seed so that the only difference between two reported numbers is the factor under test.

1) Data preparation. Records are loaded, the two leakage-prone variables of Section III-A are removed, the derived features are computed, and the cohort is split into a training set and a held-out test set by stratified sampling. Imputation medians and standardisation statistics are estimated on the training split alone and then applied to the test split, so no test information reaches the model.

2) Partitioning. The training split is divided among K clients either by Dirichlet sampling at a chosen concentration or by the natural enrolment-year variable. The total variation distance of (1) is recorded for the partition actually drawn, not for the nominal concentration, because the realised skew of a draw and its expectation can differ appreciably at small K.

3) Initialisation. The server draws one parameter vector and broadcasts the same vector to every client, so that all strategies within a seed start from an identical point and no difference between them can be attributed to initialisation.

4) Local training. Each selected client runs the required number of local epochs. In the private setting the local update is the DP-SGD step of Section III-D; otherwise it is plain mini-batch stochastic gradient descent.

5) Aggregation. The server combines the returned parameters using the rule under test: the weighted mean, the proximal-corrected mean, server momentum, the coordinate-wise median, the trimmed mean or Krum.

6) Privacy accounting. The accountant is advanced by the number of local steps actually taken at each client, and the budget reported for the round is the maximum over clients rather than the mean.

7) Evaluation. AUROC, AUPRC, subgroup AUROC and the membership inference AUC are computed on the untouched test split at the final round.

Steps three to five repeat for sixty communication rounds, and the whole sequence repeats for five independent seeds, giving the 768 training runs reported in this paper.

The implementation is written in Python. The network, the per-example gradients, the clipping operation and the Gaussian mechanism are implemented directly from the closed-form expressions for the architecture of Section III-C using NumPy, rather than taken from a privacy library. This is a deliberate choice: it keeps the sensitivity argument explicit and auditable, and it removes the risk that an unexamined library default silently changes the guarantee being reported. Scikit-learn supplies the stratified splits and the AUROC and AUPRC computations, SciPy supplies the paired significance tests, and Matplotlib produces the figures. The Rényi accountant follows the analytical moments formulation of Wang et al. [15]. Because the model has only 481 parameters, the entire grid of 768 runs completes on a single CPU workstation without a graphics processor, which keeps the study reproducible on ordinary departmental hardware.

Statistical analysis follows the same protocol throughout. Every reported value is the mean over the independent seeds and is quoted with its standard deviation. Two strategies are compared with a paired t-test over matched seeds, which is valid here because both strategies see the same partition and the same initialisation within a seed; the pairing removes the partition-to-partition variance and is the reason a difference of 4.68 AUROC points reaches $p = 9 \times 10^{-5}$ on only five runs. Differences smaller than the pooled seed standard deviation are reported as statistically indistinguishable rather than as an ordering, and the sensitivity results of Table III use three seeds because they are read as trends rather than as tested hypotheses.

G. Alternative approaches considered

Robustness of a measurement depends as much on the alternatives rejected as on the method adopted, so the choice made at each design point is recorded here.

1) Aggregation rule. FedAvgM and FedBN were candidates alongside FedAvg and FedProx. FedAvgM was implemented and is reported in Table I. FedBN was not run because the network carries no normalisation layers for a client to retain [21], and the benchmark of Manthe et al. indicates that elaborate personalisation adds little over plain averaging at this problem size [22].

2) Privacy mechanism. Central differential privacy applied at the server was rejected because it requires trusting the aggregator with clean updates, which is precisely the assumption this study avoids. Secure aggregation and homomorphic encryption were considered and are discussed in Section II; they protect the individual update rather than the released model and leave the learning dynamics unchanged, so they cannot be placed on the accuracy axis measured here.

 DOI : 10.35291/2454-9150.2026.0044

3) Privacy accounting. The Rényi accountant was preferred to basic composition and to the earlier moments accountant because it gives the tightest budget for the Poisson sub-sampled Gaussian mechanism at the composition depth used here. The looser conventions available in the literature are exactly what makes cross-paper comparison unreliable, which is why the convention is stated in full in Section III-D.

4) Heterogeneity model. Dirichlet partitioning is the standard synthetic device, but it was paired with the natural enrolment-year split so that the synthetic curves could be calibrated against one federation that was not constructed by the experimenter.

5) Robust aggregation. The coordinate-wise median, the trimmed mean and Krum were all implemented and compared in Fig. 4, rather than adopting a single defence in advance and reporting only its behaviour.

6) Missing data. Median imputation with an explicit missingness indicator was preferred to dropping the 1,350 incomplete subjects, which would have discarded 17 per cent of the cohort and altered the outcome rate, and to model-based imputation, which would itself have to be fitted federatively to remain faithful to the setting.

IV. RESULTS AND DISCUSSION

A. Collaboration is worth more than federation costs

Table I gives the central comparison between the two reference points and the four aggregation strategies. Hospitals training alone reached an AUROC of 0.7851, while federated averaging over the same hospitals reached 0.8319, a gain of 4.68 points that is significant across seeds (paired $t = 16.0$, $p = 9 \times 10^{-5}$). Pooling all records centrally, which privacy law does not permit, would give 0.8389. Federation therefore keeps 99.2 per cent of the pooled-data accuracy, and the remaining loss of 0.70 points is about one seventh of the gain from collaborating. This agrees with the 99 per cent model quality reported across ten institutions by Sheller et al. [4].

Table I. Comparison of aggregation strategies on FLCHAIN (ten clients, five seeds). Values are mean ± standard deviation across seeds.

Method	AUROC	AUPRC	Rounds to 0.82
Centralised (pooled)	0.8389 ± 0.0012	0.7034	1.0
Isolated local training	0.7851 ± 0.0091	0.6161	never
FedAvg	0.8319 ± 0.0045	0.6901	2.6
FedAvgM ($\beta = 0.9$)	0.8158 ± 0.0131	0.6670	4.4
FedProx ($\mu = 0.1$)	0.8339 ± 0.0048	0.6921	3.8
FedProx ($\mu = 1.0$)	0.8355 ± 0.0031	0.6926	13.6

Two further outcomes are visible in Table I. The AUPRC column moves in the same direction as the AUROC but with a wider spread between the best and the worst strategy, 0.7034 against 0.6161, which confirms that the imbalanced positive class is where the differences between strategies are concentrated. The final column reports the number of rounds needed to reach an AUROC of 0.82 and shows that the cost of a strategy is not only its final accuracy: FedProx at a coefficient of 1.0 attains the best accuracy in the table but needs 13.6 rounds against 2.6 for federated averaging, which is a five-fold increase in communication for a gain of 0.36 points. Isolated training never reaches the threshold at all. For a consortium paying for every communication round, FedProx at a coefficient of 0.1 is therefore the better operating point of the two.

B. Heterogeneity matters less than expected

Moving from an identically distributed split (TV = 0.02) to the most extreme partition tested (TV = 0.46, where several clients hold only one outcome class) cost federated averaging just 1.36 points, from 0.8406 to 0.8270. Even at its worst the federated model stayed 4.2 points above isolated training. Heterogeneity is usually presented as the main obstacle to federated learning in healthcare, but on a moderately sized tabular task it is a second-order effect, an order of magnitude smaller than the cost of privacy reported next.

The proximal correction behaved in an unexpected way. In the moderate range it helped, giving 0.8355 against 0.8319 at alpha = 0.5, but at alpha = 0.05 the ordering reversed to 0.8159 against 0.8270. The reason is that under extreme skew some clients hold only one class, and a strong proximal penalty holds such a client so close to the broadcast parameters that it contributes almost nothing. The coefficient must therefore be tuned against the measured heterogeneity rather than fixed in advance.

C. The cost of differential privacy

Fig. 2 and Table II report the same experiment from two directions: the figure shows the shape of the trade-off across strategies, and the table gives the underlying values together with the noise multiplier that produces each budget. The curve has a clear knee. At a budget of 20 the model keeps 98.8 per cent of its non-private accuracy and at 10 it keeps 96.4 per cent, but below that it falls quickly, to 86.5 per cent at 2 and 79.3 per cent at 1. The AUPRC falls faster than the AUROC throughout, which matters clinically because the noise damages the ranking of the high-risk minority group most, and that is the group a mortality model exists to find. At a budget of 1 the AUPRC has dropped from 0.690 to 0.427, a relative loss of 38 per cent against 21 per cent for AUROC.

 DOI : 10.35291/2454-9150.2026.0044

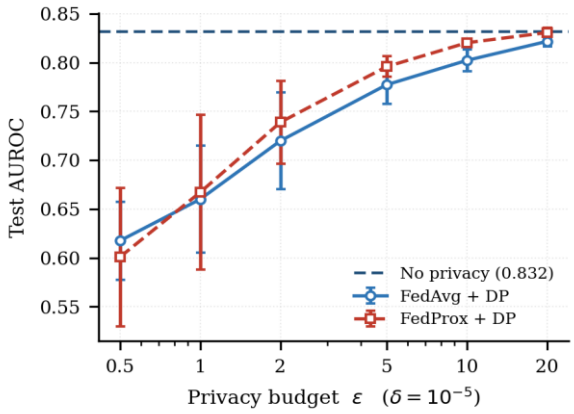

Fig. 2. Test AUROC against the privacy budget ε at δ = 10^{-5} for FedAvg and FedProx, with the non-private reference shown as a dashed line. Error bars are one standard deviation over five seeds. FedProx retains more accuracy than FedAvg at every usable budget, and the separation between the two curves widens as the budget tightens.

Two features of Fig. 2 carry the argument. The first is the knee near ε = 5: to the right of it the curves are flat and the privacy guarantee is nearly free, while to the left of it accuracy falls steeply and the error bars widen, so a consortium operating below that point pays twice, once in the mean and once in the run-to-run reliability of the model it deploys. The second is that the FedProx curve lies above the FedAvg curve everywhere the model is usable, which is the observation developed in Section IV-D.

Table II. Privacy and accuracy trade-off on FLCHAIN under FedAvg (δ = 10^{-5}, five seeds). σ is the noise multiplier required to achieve the stated budget.

ε	σ	FedAvg AUROC	% of non-private
0.5	66.74	0.6174 ± 0.0398	74.2
1	35.24	0.6600 ± 0.0546	79.3
2	18.75	0.7198 ± 0.0496	86.5
5	8.36	0.7772 ± 0.0193	93.4
10	4.73	0.8021 ± 0.0110	96.4
20	2.82	0.8216 ± 0.0054	98.8
No privacy	0	0.8319 ± 0.0045	100

The σ column of Table II explains the shape of the curve and is the practical outcome of this section. The noise multiplier is not linear in the budget: halving ε from 20 to 10 raises σ from 2.82 to 4.73, but halving it again from 1 to 0.5 raises σ from 35.24 to 66.74. Each further halving of the budget therefore costs progressively more noise per parameter, which is why the accuracy column collapses rather than declines. The standard deviations behave the same way, rising from 0.0054 at ε = 20 to 0.0398 at ε = 0.5, so at strict budgets the model that a hospital happens to obtain from one training run becomes an unreliable guide to the model it would obtain from another.

D. A proximal term halves the privacy budget

FedProx combined with private training beat federated averaging at every budget from 2 upwards, by 0.0186 at a budget of 5 and 0.0180 at 10. This is about five times the advantage of 0.0036 that the same term gives without noise. The reason is that the proximal penalty acts as a variance reducer once the update is dominated by injected noise: it holds the local model steady and averaging then removes the rest.

The practical consequence is direct. FedProx with privacy at a budget of 10 gave 0.8202, and federated averaging at a budget of 20 gave 0.8216. The two are statistically indistinguishable, so using the proximal term halves the privacy budget at no measurable cost in accuracy, for a one-line change in the local objective. A second free lever is the clipping norm: at a fixed budget of 10, reducing it from 4.0 to 0.25 improved AUROC from 0.7431 to 0.8302.

E. Privacy is not equally fair to all patients

Fig. 3 shows performance split by age band, with the pooled result plotted alongside the oldest and youngest thirds of the test set. Between no privacy and a budget of 10, overall AUROC fell from 0.8319 to 0.8021, a relative loss of 3.6 per cent, while the youngest third fell from 0.6561 to 0.6071, a relative loss of 7.5 per cent. The weakest subgroup therefore degrades about 2.1 times faster than the average. At a budget of 1 the youngest band reached 0.5073, which is no better than guessing, even though the overall figure of 0.6600 still looks like a weak but working model.

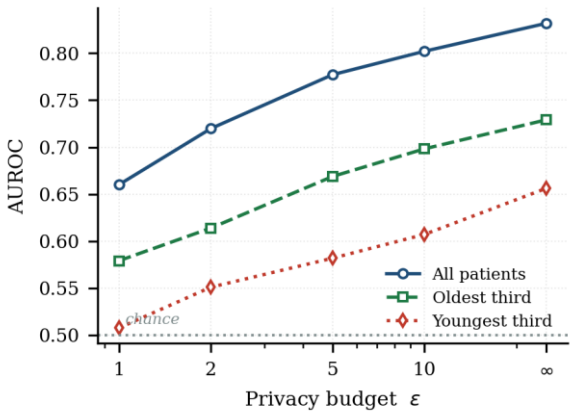

Fig. 3. Subgroup AUROC against the privacy budget ε for all patients, the oldest third and the youngest third of the test cohort. The dotted line marks chance performance. The three curves diverge as the budget tightens, showing that differential privacy harms the weakest patient subgroup fastest.

The outcome to take from Fig. 3 is the divergence of the three curves rather than the level of any one of them. The vertical gap between the pooled curve and the youngest-third curve widens monotonically as the budget tightens, and at ε = 1 the youngest curve has met the chance line while the pooled curve is still well above it. An aggregate metric reported at that operating point would therefore describe a model that has stopped working entirely for one third of the patients it is

 DOI : 10.35291/2454-9150.2026.0044

meant to serve, and no aggregate figure in the study reveals this.

The mechanism is simple. A subgroup for which the model has a weaker signal contributes smaller gradients, and a fixed amount of noise removes a larger share of a weaker signal. Any group that is harder to predict, or under-represented in training, is hit harder by the same mechanism. This confirms with concrete numbers the concern raised in the review literature [11], and argues that subgroup results should be reported whenever differential privacy is applied to a clinical model. The flexible fairness regularisation of Xing et al. [26] offers one route to correcting the disparity, although it has not yet been evaluated under a privacy budget.

F. Measured leakage is far below the formal bound

With no privacy mechanism at all, the membership inference attack achieved an AUC of only 0.5101 with a standard deviation of 0.0009. This is measurably above chance, but far below the 60 to 80 per cent success rates reported against deep networks on brain images [19]. Applying privacy reduced the attack only slightly, to 0.5034 at a budget of 0.5. In other words, moving from no privacy to a very strict budget cut measured leakage by 0.0067 while cutting accuracy by 0.2145, a factor of thirty larger.

Three points qualify this. The low leakage reflects a model that barely overfits, since 481 parameters fitted to 5,905 records produce training and test loss distributions that overlap almost completely, and inference success is known to track overfitting [7]. The loss-threshold attack is a baseline, and stronger attacks exist; the tighter black-box auditing procedure of Annamalai and De Cristofaro recovers a substantially larger empirical epsilon than earlier audits on image data [27], so the figure reported here should be read as a lower bound on realised leakage rather than as its true value. Leakage is also known to be weaker on tabular data than on high-dimensional data. The conclusion is not that privacy is unnecessary, but that the budget should be chosen against a measured threat model: at a budget of 10 a meaningful guarantee costs 3.6 per cent of accuracy, whereas pushing to 1 gives up a fifth of it to close a channel that measurement suggests was barely open.

G. Robustness to dishonest hospitals

All the results above assume every hospital follows the protocol. Fig. 4 removes that assumption and compares four aggregation rules across three attack scenarios. Plain averaging is fragile: two malicious clients out of ten using a sign-flipping attack reduce it to an AUROC of 0.5001, which is exactly random, and Gaussian corruption reduces it to 0.4589. The coordinate-wise median is the most reliable defence, costing nothing when there is no attack and holding 0.8252 under two attackers. Krum resists sign flipping but collapses to 0.5084 against free riders, because identical free-rider updates form the tightest cluster it looks for. A robust aggregation rule is therefore not optional in a real deployment.

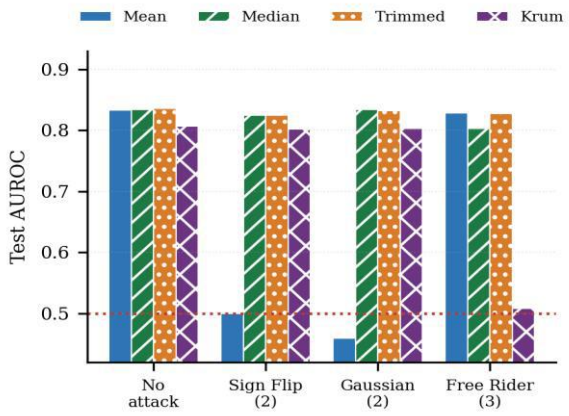

Fig. 4. Test AUROC of four aggregation rules under three attack scenarios, with the number of malicious clients out of ten given in brackets. The dotted line marks chance performance. Plain averaging fails under sign-flipping and Gaussian corruption; the coordinate-wise median and the trimmed mean do not.

Read across the groups of Fig. 4, the outcome is a clear ordering. The mean is the only rule that ever falls to chance, and it does so in two of the three attacks. The median and the trimmed mean hold their no-attack performance in every scenario, so their robustness is obtained at no measurable cost in the honest case, which removes the usual argument for deferring a defence until an attack is observed. Krum is the exception that defines the limit of the comparison: it is the only rule whose failure is caused by the attack exploiting its selection criterion rather than by the magnitude of the corruption, which is why a defence should be chosen against a stated threat model rather than by reputation.

H. Which settings actually matter

Table III reports how sensitive the system is to its main configuration choices, which is useful for anyone reproducing the work. Only three settings matter. Client participation must be at least 40 per cent; at 20 per cent the mean drops and the variation across seeds rises about fourfold, because too few clients contribute in each round for the average to be stable. The number of local epochs shows a steady penalty, from 0.8366 with one epoch to 0.8238 with eight, since more local work means more drift before averaging. The clipping norm, discussed in Section IV-D, is the most important setting once privacy is switched on. Batch size, learning rate and network width made almost no difference; a network with eight hidden units performed as well as one with sixty-four, which suggests the task is limited by the information in the features rather than by model capacity.

 DOI : 10.35291/2454-9150.2026.0044

Table III. Sensitivity of the federated model to its main configuration settings (three seeds). Values are mean AUROC ± standard deviation.

Setting	Value	AUROC
Client participation	0.2	0.8249 ± 0.0096
	0.4	0.8342 ± 0.0037
	1.0	0.8348 ± 0.0026
Local epochs	1	0.8366 ± 0.0024
	4	0.8311 ± 0.0025
	8	0.8238 ± 0.0042
Hidden units	8	0.8361 ± 0.0020
	64	0.8353 ± 0.0017

The practical outcome of Table III is that the reproduction burden is small. The spread across the whole sweep is 0.0128 AUROC, which is smaller than the 0.0298 cost of moving to a privacy budget of 10 and far smaller than the 4.68-point gain from federating at all, so a group repeating this work does not need to match the hyper-parameters exactly to recover the same conclusions. It need only keep client participation above 40 per cent, keep local epochs low, and tune the clipping norm once privacy is enabled.

I. External check and limitations

On the Breast Cancer Wisconsin data with five clients, centralised training reached 0.9594 accuracy, federated averaging 0.9580 and isolated training 0.8932, which confirms the implementation is sound. Under privacy at a budget of 1.9 our federation reached 0.8741, whereas Shukla et al. report 0.961 at the same nominal budget [9]. We do not read this as an error in either study. The likely cause is the accounting convention, since our budget is per record, worst case across clients, and composed over every local step, and each of those choices is the conservative one.

The study has clear limits. FLCHAIN is an epidemiological cohort rather than a hospital record extract, and its thirteen predictors are fewer than an intensive care record offers. The network is small compared with clinical imaging models, and both heterogeneity effects and leakage are expected to grow with model size. Only one attack was used in the audit, and only record-level privacy was evaluated.

V. CONCLUSION AND FUTURE SCOPE

This paper set out to measure, rather than assume, what federated learning and differential privacy actually cost in clinical risk prediction. The investigation used two real clinical datasets, ten simulated hospitals and one non-synthetic federation built from enrolment year, and it reports 768 seeded training runs under a single privacy convention stated in full: a record-level guarantee, taken as the worst case across clients and composed over every local step at $\delta = 10^{-5}$. Five findings follow from that investigation.

First, collaboration is worth considerably more than federation costs. Federated averaging reached an AUROC of 0.8319 against 0.8389 for centrally pooled data and 0.7851 for hospitals training alone, so federation retains 99.2 per cent of the pooled-data accuracy while the gain from collaborating is roughly seven times the loss from federating.

Second, differential privacy rather than data heterogeneity is the binding constraint. Moving from an identically distributed split to the most extreme partition tested cost 1.36 AUROC points, whereas tightening the budget from 10 to 1 cost 14.2 points. On a task of this size heterogeneity is a second-order problem.

Third, a proximal term halves the privacy budget. FedProx under privacy at $\varepsilon = 10$ gave 0.8202 and federated averaging at $\varepsilon = 20$ gave 0.8216, figures that are statistically indistinguishable, and aggressive clipping recovers most of the remaining loss, raising AUROC from 0.7431 to 0.8302 at a fixed budget of 10. Both are one-line changes to the local objective.

Fourth, the formal guarantee is conservative for this class of model. A membership inference adversary scored only 0.5101 against a completely unprotected model, so moving to $\varepsilon = 0.5$ cut measured leakage by 0.0067 while cutting accuracy by 0.2145, a factor of thirty larger. The budget should therefore be selected against a measured threat model rather than against convention.

Fifth, privacy noise is not equally fair. Between no privacy and $\varepsilon = 10$ the overall AUROC fell by 3.6 per cent while the youngest third fell by 7.5 per cent, a ratio of about 2.1, and at $\varepsilon = 1$ that subgroup performed no better than chance while the aggregate figure still resembled a working model. Subgroup results should therefore be reported whenever differential privacy is applied to a clinical model.

Taken together these findings support a concrete recommendation for a hospital consortium: adopt federated averaging with a small proximal term and a tight clipping norm, operate near $\varepsilon = 10$ rather than at the stricter budgets often quoted, defend the aggregation step with the coordinate-wise median, and report subgroup performance alongside the headline metric. The findings hold within the limits stated in Section IV-I, namely an epidemiological cohort with thirteen predictors, a compact network, a single attack in the audit and a record-level guarantee only.

Future work follows directly from those limits. The protocol should be repeated on credentialed intensive care data such as MIMIC-IV, which carries real hospital identifiers, and extended to larger models where leakage is expected to rise and stronger attacks become relevant. Fairness-aware private training, for example through group-specific clipping norms or the flexible fairness regularisation of Xing et al. [26], is the obvious next step given the disparity reported here, and tighter empirical auditing along the lines of Annamalai and De Cristofaro [27] would narrow the gap between the formal budget and the leakage a hospital actually faces. The complete implementation and a reproduction notebook are available from the corresponding author.

 DOI : 10.35291/2454-9150.2026.0044

ACKNOWLEDGMENT

The authors thank the Department of Computer Science & Engineering for the computing facilities provided, and acknowledge the investigators of the Mayo Clinic Olmsted County study and the University of Wisconsin for making their clinical datasets openly available for research. The authors also thank the anonymous reviewers, whose comments on the methodology, the presentation of the figures and tables, and the concluding remarks improved this paper.

REFERENCES

[1] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated learning for healthcare informatics," J. Healthc. Inform. Res., vol. 5, no. 1, pp. 1–19, 2021.

[2] R. S. Antunes, C. A. da Costa, A. Küderle, I. A. Yari, and B. Eskofier, "Federated learning for healthcare: Systematic review and architecture proposal," ACM Trans. Intell. Syst. Technol., vol. 13, no. 4, pp. 1–23, 2022.

[3] I. Dayan, H. R. Roth, A. Zhong, et al., "Federated learning for predicting clinical outcomes in patients with COVID-19," Nat. Med., vol. 27, no. 10, pp. 1735–1743, 2021.

[4] M. J. Sheller, B. Edwards, G. A. Reina, et al., "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," Sci. Rep., vol. 10, art. 12598, 2020.

[5] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artif. Intell. Statist., 2017, pp. 1273–1282.

[6] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Symp. Secur. Privacy, 2017, pp. 3–18.

[7] H. Hu, Z. Salcic, L. Sun, G. Dobbie, and X. Zhang, "Source inference attacks in federated learning," in Proc. IEEE Int. Conf. Data Mining, 2021, pp. 1102–1107.

[8] K. Wei, J. Li, M. Ding, et al., "Federated learning with differential privacy: Algorithms and performance analysis," IEEE Trans. Inf. Forensics Security, vol. 15, pp. 3454–3469, 2020.

[9] S. Shukla, S. Rajkumar, A. Sinha, M. Esha, K. Elango, and V. Sampath, "Federated learning with differential privacy for breast cancer diagnosis enabling secure data sharing and model integrity," Sci. Rep., vol. 15, art. 13061, 2025.

[10] R. Tertulino, "A robust pipeline for differentially private federated learning on imbalanced clinical data using SMOTETomek and FedProx," arXiv:2508.10017, 2025.

[11] M. Mohammadi, M. Vejdanihemmat, M. Lotfinia, M. Rusu, D. Truhn, A. Maier, and S. Tayebi Arasteh, "Differential privacy for medical deep learning: Methods, tradeoffs, and deployment implications," npj Digit. Med., vol. 9, no. 1, art. 93, 2026.

[12] X. Li, K. Huang, W. Yang, S. Wang, and Z. Zhang, "On the convergence of FedAvg on non-IID data," in Proc. Int. Conf. Learn. Represent., 2020.

[13] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in Proc. Mach. Learn. Syst., vol. 2, 2020, pp. 429–450.

[14] A. Ziller, D. Usynin, R. Braren, M. Makowski, D. Rueckert, and G. Kaissis, "Medical imaging deep learning with differential privacy," Sci. Rep., vol. 11, art. 13524, 2021.

[15] Y.-X. Wang, B. Balle, and S. Kasiviswanathan, "Subsampled Rényi differential privacy and analytical moments accountant," in Proc. 22nd Int. Conf. Artif. Intell. Statist., 2019, pp. 1226–1235.

[16] L. Bai, H. Hu, Q. Ye, H. Li, L. Wang, and J. Xu, "Membership inference attacks and defenses in federated learning: A survey," ACM Comput. Surv., vol. 57, no. 4, pp. 1–35, 2024.

[17] A. Dispenzieri, J. A. Katzmann, R. A. Kyle, et al., "Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population," Mayo Clin. Proc., vol. 87, no. 6, pp. 517–523, 2012.

[18] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, "Nuclear feature extraction for breast tumor diagnosis," in Proc. IS&T/SPIE Int. Symp. Electron. Imaging, vol. 1905, 1993, pp. 861–870.

[19] U. Gupta, D. Stripelis, P. K. Lam, P. Thompson, J. L. Ambite, and G. Ver Steeg, "Membership inference attacks on deep regression models for neuroimaging," in Proc. Med. Imaging Deep Learn., 2021, pp. 228–251.

[20] E. H. Lee, M. Han, J. Wright, et al., "An international study presenting a federated learning AI platform for pediatric brain tumors," Nat. Commun., vol. 15, art. 7615, 2024.

[21] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "FedBN: Federated learning on non-IID features via local batch normalization," in Proc. Int. Conf. Learn. Represent., 2021.

[22] M. Manthe, S. Duffner, and C. Lartizien, "Federated brain tumor segmentation: An extensive benchmark," Med. Image Anal., vol. 97, art. 103270, 2024.

[23] M. Firdaus, S. Noh, Z. Qian, H. T. Larasati, and K.-H. Rhee, "Blockchain-based federated learning with homomorphic encryption for privacy-preserving healthcare data sharing," Internet of Things, vol. 31, art. 101579, 2025.

[24] D. Stripelis, U. Gupta, H. Saleem, et al., "A federated learning architecture for secure and private neuroimaging analysis," Patterns, vol. 5, no. 8, art. 101031, 2024.

[25] M. R. Hasan, M. I. Ahmed, S. Saha, T. K. Ishika, H. A. Shoaib, and M. J. Hossen, "Multi-modal federated learning with differential privacy for privacy-preserving healthcare AI," Sci. Rep., vol. 16, art. 21253, 2026.

[26] H. Xing, R. Sun, J. Ren, J. Wei, C.-M. Feng, X. Ding, et al., "Achieving flexible fairness metrics in federated medical imaging," Nat. Commun., vol. 16, art. 3342, 2025.

[27] M. S. M. S. Annamalai and E. De Cristofaro, "Nearly tight black-box auditing of differentially private machine learning," in Proc. 38th Conf. Neural Inf. Process. Syst. (NeurIPS), 2024.

 DOI : 10.35291/2454-9150.2026.0044