

Original Research

Received 27 June 2026

Revised 28 July 2026

Accepted 10 Aug 2026

Natural Resources for Human
Health 2026, Vol. 6 (17s): 1452–1461

KEYWORDS:

Federated learning; differential
privacy; model calibration;
membership inference;
cardiovascular risk.

Privacy Noise Damages Calibration in Federated Clinical Models

Abdul Ahad^{*1,2}, Mohammed Ali Shaik¹, Md Riyazuddin³, N
Udaya Bhaskara Varma⁴, Md Jaffar Sadiq⁵

¹ School of Computer Science and Artificial Intelligence, SR University, Warangal
506371, Telangana, India

² School of Engineering, Anurag University, Hyderabad 500088, Telangana, India

³ Dept of CSE-AIML, Sreenidhi University, Hyderabad 501301, Telangana, India

⁴ Department of Mathematics, SRKR Engineering College, Bhimavaram 534204, Andhra
Pradesh, India

⁵ Department of CSE-DS, Sreenidhi Institute of Science and Technology, Hyderabad
501301, Telangana, India

ABSTRACT: Hospitals cannot pool patient records, and federated learning is widely proposed as the remedy. Most evidence comes from simulated federations cut from one dataset, which understates the difficulty because every shard shares one measurement process. We instead used a cardiology cohort of 920 patients collected independently at four hospitals in the United States, Hungary and Switzerland, where coronary disease prevalence ranges from 36.1 to 93.5 percent and missing data from 0.2 to 28.8 percent. Local training, pooled training, federated averaging, federated averaging with a proximal term, and federated averaging under record level differential privacy were compared over 25 repeated stratified runs, under one protocol that scores each hospital on its own held out records and a second that scores the shared model on a hospital it was never trained on. A membership inference attack measured leakage directly. Federated averaging recovered 87 percent of the benefit of pooling (area under the curve 0.798 against 0.800 pooled and 0.785 local), but that benefit was almost entirely confined to Long Beach VA, which gained 0.050 while no other site gained more than 0.001. The proximal term changed nothing. Privacy damaged calibration before ranking: at a budget of 8 the calibration error rose 14 percent while the area under the curve fell 1.2 percent. Privacy cost 1.6 to 2.7 times more on an unseen hospital, and federated training leaked less than local training (0.5048 against 0.5196). Federations should report per site outcomes and calibration, not averages and discrimination alone.

1. INTRODUCTION

Machine learning for clinical risk needs more patients than any one hospital holds. It also needs records that a hospital is usually not allowed to send anywhere. Federated learning was proposed to resolve this by moving the model to the data instead of the data to the model (McMahan et al., 2017). Each site trains on its own records and returns only model weights, which a coordinating server averages (Kairouz et al., 2021). Health care has taken the idea up quickly, and there are now convincing demonstrations at scale in imaging and in critical care (Rieke et al., 2020; Sheller et al., 2020; Dayan et al., 2021), in decentralised training that dispenses with the central server altogether (Warnat-Herresthal et al., 2021), and in predicting mortality from the records of several hospitals in one health system (Vaid et al., 2021).

Most published evaluations, however, build their federation by taking a single dataset and dividing it among artificial clients. That is convenient and it is reproducible, but it understates the problem in a specific way. Every shard is drawn from one measurement process, one recording convention and one patient population, so the sites differ only in which rows they happen

to hold. Real hospitals differ in what they record, how completely they record it, and which patients walk through the door (Li et al., 2020a; Zhao et al., 2018). The studies that do use hospitals which were genuinely separate tend to report one pooled figure of merit for the federation (Brisimi et al., 2018), and reviews of the area keep asking for evaluations that sit closer to deployment (Xu et al., 2021; Kaissis et al., 2020). A method that survives a synthetic split has not yet been shown to survive that.

Two questions matter to a hospital deciding whether to join a federation, and neither is usually answered directly. The first is who gains. An average across sites can look convincing while hiding the fact that one participant gained everything and the others gained nothing, and a hospital cannot act on an average that may not describe it. The second is what a formal privacy guarantee costs. Federated learning by itself does not provide one, because the weights a site returns are still a function of its patients (Shokri et al., 2017; Nasr et al., 2019). Differential privacy supplies the guarantee (Dwork and Roth, 2014; Abadi et al., 2016), and it can be combined with secure aggregation inside a federation (Truex et al., 2019); the cost is usually reported as a fall in the area under the curve, which is only part of the story for a model whose output is meant to be read as a probability.

This study addresses both questions on a cohort that was genuinely collected at four separate hospitals, not split into four. Our contributions are as follows. We measure who benefits from federation site by site rather than on average, and find the benefit is concentrated almost entirely in one hospital. We report a clear null result for a widely recommended correction for heterogeneous data. We show that differential privacy damages calibration well before it damages ranking, which matters because a clinical threshold is applied to a probability. We show the privacy cost is roughly twice as large when the model is used at a hospital outside the training group. Finally we measure leakage with an attack rather than assuming it, and find that federation itself reduces leakage relative to training alone.

2. MATERIALS AND METHODS

2.1 The federation

We used the cardiology cohort assembled by Detrano and colleagues (1989), which is held as four separate files because it was collected as four separate studies: the Cleveland Clinic Foundation, the Hungarian Institute of Cardiology in Budapest, two hospitals in Switzerland, and the Long Beach Veterans Affairs Medical Center. The sites are treated here as four clients that never pool records. Each patient has thirteen clinical and demographic attributes, and the outcome is the presence of coronary artery disease on angiography, coded as any narrowing above fifty percent in a major vessel.

Table 1. The four participating hospitals. The sites disagree on disease prevalence, on how much of the record is filled in, and on the age of the patients they see.

Site	Records	With disease	Prevalence (%)	Missing values (%)	Mean age (years)	Male (%)
Cleveland	303	139	45.9	0.2	54.4	68.0
Hungary	294	106	36.1	20.5	47.8	72.4
Switzerland	123	115	93.5	24.8	55.3	91.9
Long Beach VA	200	149	74.5	28.8	59.4	97.0
All sites	920	509				

The disagreement between these sites is not a modelling nuisance, it is the object of study. As **Table 1** shows, prevalence runs from 36.1 percent at Hungary to 93.5 percent at Switzerland. Missing values run from 0.2 percent at Cleveland to 28.8 percent at Long Beach VA, and serum cholesterol is not recorded at all in the Switzerland file. Mean age differs by 11.5 years across sites. A federation of this shape is the hard case, and it is the case a synthetic split cannot produce.

2.2 Preparing data without leaving the site

Two rules were enforced so that a federated experiment could not quietly become a centralised one. First, missing values were filled using statistics computed inside a site from that site's own training rows, never from the federation. A missing indicator was added for every attribute, so a model can tell a filled value from a measured one, which matters when a whole column is absent at one site. Second, feature scaling used federated statistics: each site contributed only a count, a sum and a sum of squares per attribute, which is the same information a secure aggregation round would release, and the shared mean and standard deviation were reconstructed from those. Values of zero recorded for cholesterol and resting blood pressure were treated as missing rather than as measurements, because a cholesterol of zero is a recording convention and not a reading.

2.3 Training

All arms used regularised logistic regression, so that a difference between arms is a difference in how the data was used and not a difference in model capacity. Five arms were compared. Local training used one site's records only. Pooled training combined all records in one place, which is not permitted in practice and is included as a ceiling. Federated averaging ran 50 rounds (**Figure 1**), each site performing one local epoch before the server took an average weighted by the number of records at each site (McMahan et al., 2017). The proximal variant added a term pulling each local solution back towards the shared one, which is the standard recommendation for heterogeneous clients (Li et al., 2020b), and was run at four strengths. The private arm applied record level differential privacy inside each site. Every arm shares the same estimator and the same preprocessing code (Pedregosa et al., 2011), so two arms differ only where the design says they differ.

2.4 Privacy

Per record gradients were clipped to a norm of 1.0 and Gaussian noise was added before averaging, following the standard construction for private stochastic gradient descent (Abadi et al., 2016). The noise scale for a target budget was obtained by Rényi accounting (Mironov, 2017) at a failure probability of 10^{-5} , giving noise multipliers from 3.06 at the loosest budget to 35.62 at the tightest. The unit protected is the individual patient record. Client level privacy, where the unit protected is a whole hospital, was considered and rejected: with four clients the average is a quarter of any one update, so the noise needed to conceal a hospital removes the signal (Geyer et al., 2017). Record level privacy is both what a four hospital federation can run and what the patient cares about.

A privacy claim resting only on the fact that no records were exchanged is weak, because the trained model is itself a function of the records. We therefore also ran a membership inference attack of the loss threshold form (Yeom et al., 2018), which needs no shadow models and is therefore difficult to tune in the defender's favour. The attacker sees a record and the model and decides whether that record was in training. Attack success is reported as the area under the curve of that decision, where 0.5 means the attacker learned nothing (Shokri et al., 2017). Stronger attacks exist which calibrate the decision record by record (Carlini et al., 2022), so the number reported here should be read as a floor on what a determined attacker could manage, not a ceiling.

2.5 Evaluation

Two protocols were used because they answer different questions. Under protocol A, every hospital divided its own records into training and test parts, the shared model was trained on the four training parts and scored on each hospital's own test part. This asks whether a hospital is better off joining than working alone. Under protocol B, the model was trained on three hospitals and scored on the fourth, whose records were never seen. This asks whether the shared model transfers, which is the harder question and the one that matters if a federation intends to publish its model.

Both protocols used 5 fold stratified cross validation repeated 5 times, giving 25 matched runs. Discrimination is reported as the area under the receiver operating characteristic curve and as average precision. Calibration is reported as the Brier score and as expected calibration error over ten equal width bins (Naeini et al., 2015), because discrimination alone does not tell a clinician whether a predicted probability can be acted on (Van Calster et al., 2019; Steyerberg et al., 2010). Poor calibration is not a rare defect of exotic models: it is documented for ordinary classifiers and for modern networks alike (Guo et al., 2017; Niculescu-Mizil and Caruana, 2005). Arms were compared with the Wilcoxon signed rank test over matched folds, with a Holm correction across the family of comparisons (Holm, 1979). Rank metrics are undefined on folds in which the held out records are all of

one class, which happens at the Switzerland site, and those folds are excluded from the rank metric rather than given a substituted value.

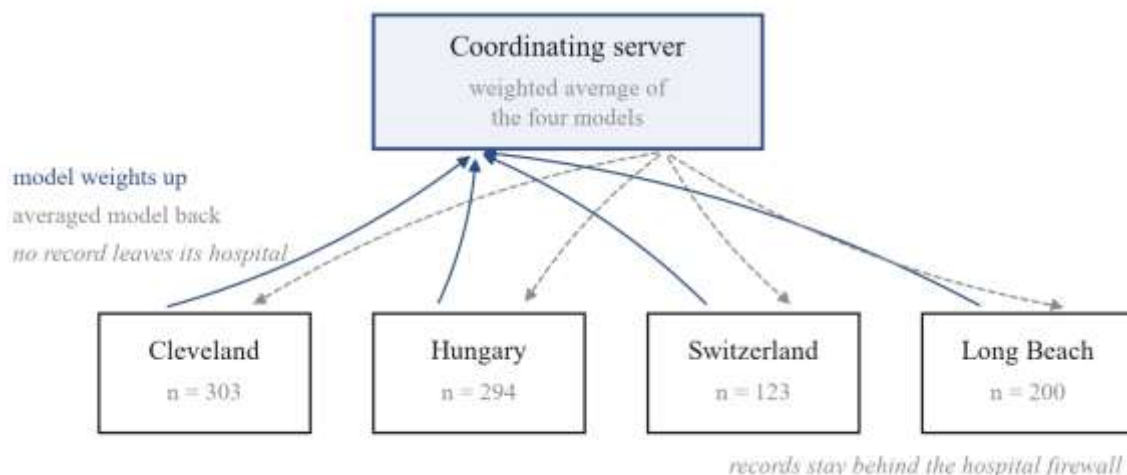


Figure 1. One federated round. Each hospital trains on its own records and returns model weights, which the coordinating server averages in proportion to the number of records held. No record leaves the hospital that collected it.

3. RESULTS AND DISCUSSION

3.1 Federated training recovers most of the benefit of pooling

Averaged across the four hospitals (**Figure 2**), local training reached an area under the curve of 0.785, pooled training reached 0.800, and federated averaging reached 0.798. Federated averaging therefore recovered 87 percent of the benefit that pooling would have given, without any record leaving its hospital. Under protocol B, where the test hospital was never trained on, federated averaging reached 0.804 against 0.800 for pooled training, so the shared model transferred at least as well as a model built on the combined data.

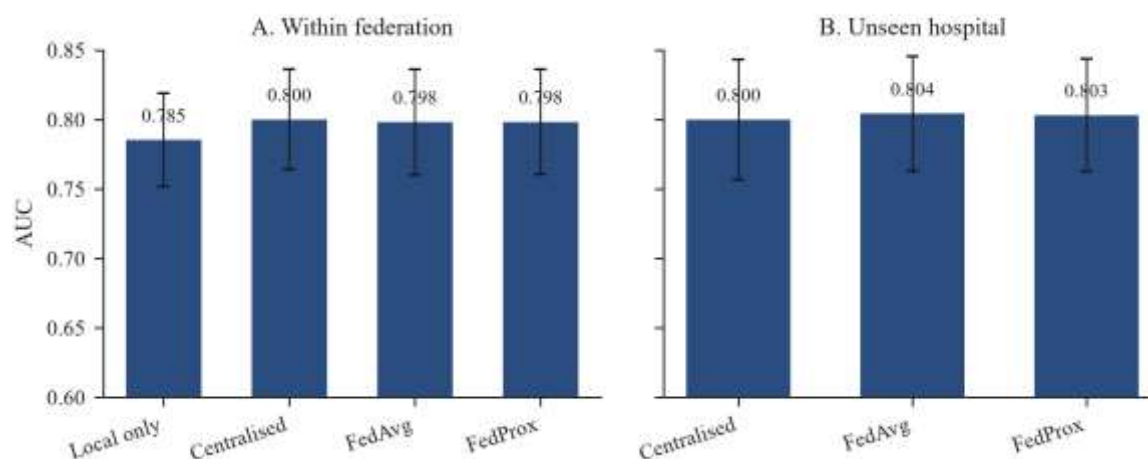


Figure 2. Discrimination by training arm under both protocols. Bars show the mean across 25 runs and whiskers show the 95 percent confidence interval of the mean.

3.2 The benefit lands almost entirely on one hospital

The average conceals the result that matters. Of the four participants (**Figure 3**), only Long Beach VA gained materially, rising from 0.696 alone to 0.747 in the federation, an increase of 0.050. This was the largest gain at any site, although it did not reach statistical significance after Holm correction (median paired difference 0.0058, $p = 0.1796$). No other hospital gained more than 0.001, and none of the other three gains approached significance. Long Beach VA is also the site with the most missing data,

at 28.8 percent, and one of the smaller ones. Notably its federated result of 0.747 exceeded the pooled result of 0.740, which suggests that averaging over sites regularises the model in a way that simply combining records does not.

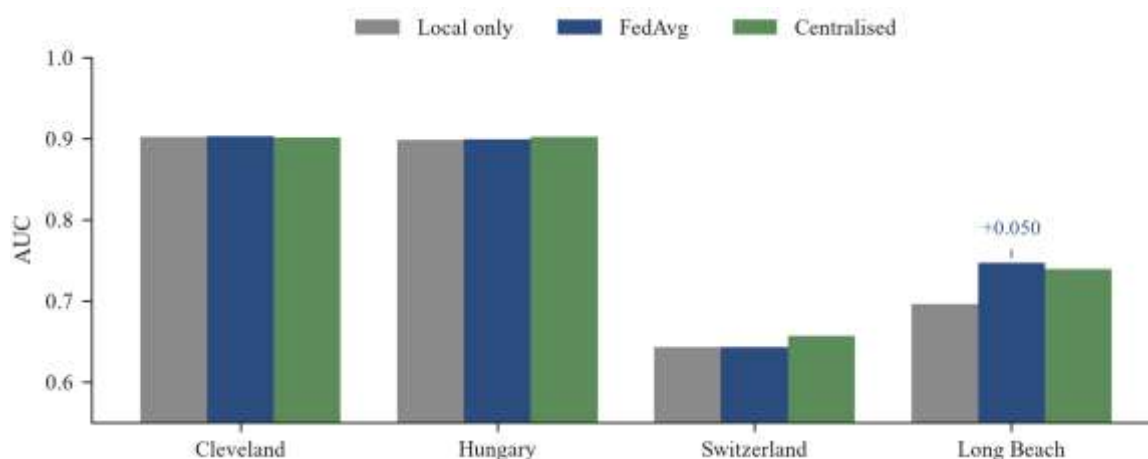


Figure 3. Discrimination at each hospital under local, federated and pooled training. The benefit of federation is concentrated at Long Beach, the site with the most missing data.

This is the practical argument for joining a federation, and it is not the argument usually made. The case is not that every participant improves. Cleveland and Hungary, which are larger and almost completely recorded, gained 0.0009 and 0.0010, which is nothing. The case is that the weakest participant improves a great deal while the strongest participants lose nothing measurable. A federation is therefore best understood as an arrangement in which strong sites donate generalisation to weak ones at no cost to themselves.

3.3 A proximal correction did not help

The proximal term is the standard recommendation when clients hold very different data (Li et al., 2020b), and this federation is about as heterogeneous as a real one gets. It made no difference. Across four settings spanning three orders of magnitude, the total spread in area under the curve was 0.00015, and no setting differed significantly from plain federated averaging. We report this as a null result rather than omitting it. With one local epoch per round the local solutions do not have time to drift far enough for the correction to have anything to pull back, and a reader planning a small federation should not expect the term to rescue them.

3.4 Privacy costs calibration before it costs ranking

The cost of privacy depends heavily on which property is measured (**Figure 4** and **Table 2**). At the loosest budget tested, the area under the curve fell by only 1.2 percent and that fall was not statistically significant, while the calibration error rose by 14 percent and that rise was significant. At the tightest budget the area under the curve fell 12.9 percent while calibration error rose 144 percent. Discrimination became significantly worse only from a budget of 4 downwards, whereas calibration was already significantly worse at 8.

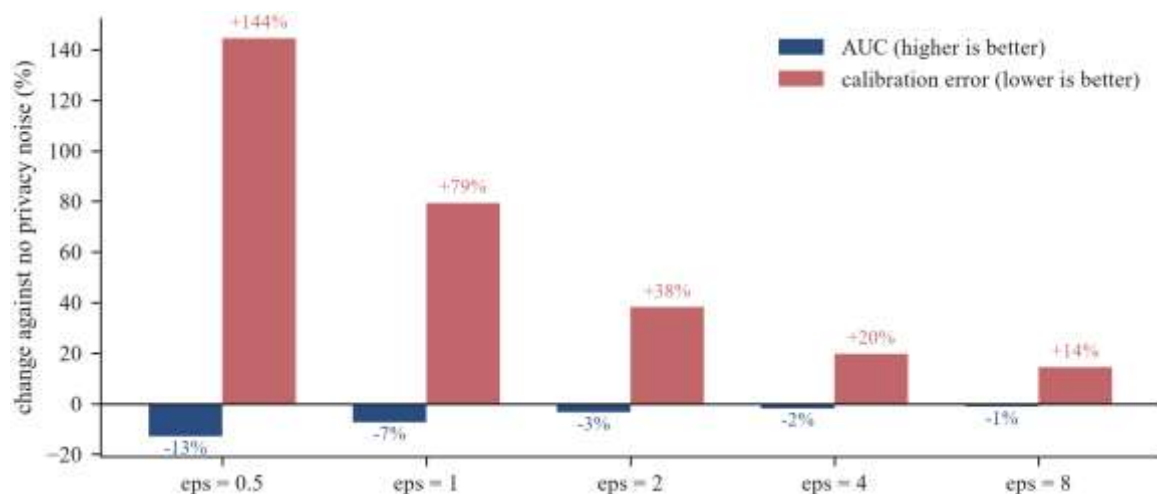


Figure 4. Change in discrimination and in calibration error under each privacy budget, relative to federated averaging with no noise. Calibration degrades several times faster than ranking.

This distinction is not academic. A model whose ranking survives but whose probabilities are inflated will still order patients correctly and will still be wrong about how many of them are at risk. Any threshold chosen on the noiseless model, and any expected number of cases derived from it, becomes incorrect. Since a private model is usually adopted precisely because it is going to be deployed, reporting only the area under the curve gives a misleading impression of what has been given up (Van Calster et al., 2019).

Table 2. Cost of each privacy budget. Discrimination and calibration are measured under protocol A. The final column is the empirical advantage of a membership inference attacker, where 0.5 means the attacker learns nothing.

Privacy budget	Noise multiplier	AUC	Change in AUC (%)	Calibration error	Change in error (%)	Attacker AUC
No noise	0	0.798	0.0	0.116	0.0	0.5048
8	3.06	0.789	-1.2	0.133	+14	0.5033
4	5.42	0.784	-1.8	0.139	+20	0.5029
2	9.96	0.771	-3.4	0.160	+38	0.5027
1	18.75	0.740	-7.3	0.208	+79	0.5022
0.5	35.62	0.695	-12.9	0.283	+144	0.5011

3.5 Privacy costs transfer about twice as much

The same noise cost markedly more when the model was applied to a hospital it had never seen (**Figure 5**). At every budget tested, the loss under protocol B was between 1.6 and 2.7 times the loss under protocol A. At a budget of 8 the model lost 0.0098 within the federation and 0.0250 on an unseen hospital. The interpretation is that noise consumes the weaker, more general signal first, and it is exactly that signal a model needs when the patient population shifts. A federation that intends to publish a private model for use elsewhere should budget for a larger loss than its internal validation suggests.

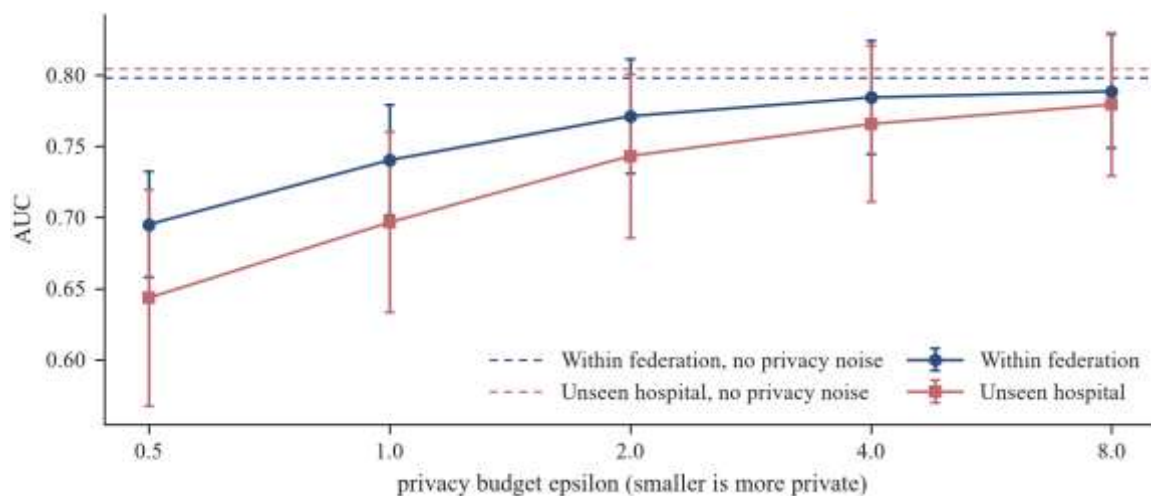


Figure 5. Discrimination against privacy budget under both protocols. The dashed lines mark federated averaging with no privacy noise. The gap widens faster for the unseen hospital.

3.6 What an attacker actually gains

Membership inference recovered very little from any of these models. Local training was the most exposed, at 0.5196, federated averaging reached 0.5048 and pooled training 0.5061, against 0.5 for an attacker who learns nothing. Adding privacy noise reduced the attacker further, to 0.5011 at the tightest budget (**Table 2**). Two things follow. First, federated training leaks less than local training here, because averaging over four sites dilutes what any one site memorised, so federation offers a modest empirical privacy benefit before any noise is added. Second, for a model of this capacity on data of this size the empirical leakage was already small, so the practical gain from a tight budget was slight while the utility cost was large.

That should not be read as an argument against differential privacy. An attack measures what one attacker achieved, whereas the guarantee bounds what any attacker could achieve, including attackers with side information that this one did not have (Nasr et al., 2019). The correct reading is that the budget should be chosen from the threat model rather than from an attack result, and that on small tabular clinical data a loose budget may be the reasonable compromise.

3.7 Communication cost

Almost all of the benefit arrived in the first round. Mean discrimination across sites after a single round was already at least as high as after 50 rounds, and further rounds did not improve it. For a convex model on a small federation the communication requirement is therefore very light, which removes one of the practical objections hospitals raise.

3.8 Discussion

Three results here are worth carrying into practice. The first is that the benefit of federation is unevenly distributed and needs to be reported site by site. A federation that reports only its average may be reporting a number that describes none of its members. In this study three of four hospitals gained essentially nothing and the fourth gained a great deal, and only a per site report makes that visible. This also suggests how to recruit participants honestly: a large well recorded hospital should be told it will probably not improve its own model, and that its contribution is to the others.

The second is that differential privacy should be reported through calibration as well as discrimination, as reporting guidance for clinical prediction models already recommends (Collins et al., 2015). The two degrade at very different rates, and the property that degrades first is the one a deployed clinical model depends on. Related work has shown that privacy noise falls unevenly on different groups within a dataset (Bagdasaryan et al., 2019). Our result is adjacent and distinct: it falls unevenly on different properties of the same model.

The third is that a privacy claim should be tested rather than asserted. Running an attack cost little and produced the useful and slightly uncomfortable finding that leakage was small before any noise was added. Reporting that honestly is more useful to a hospital than a claim that noise was applied and privacy therefore achieved.

This study has clear limitations. The cohort is small at 920 patients and comes from four sites, so the federation is at the low end of what the term usually describes, and results on a federation of dozens of hospitals may differ. A cohort of this size also sits below what current guidance would require if the aim were to develop a deployable risk model, rather than to compare training arms on matched folds as we do here (Riley et al., 2020). The data was collected some decades ago and clinical practice has changed, so the absolute performance figures should not be read as current clinical benchmarks; the comparisons between arms, which all use the same data, are what carry the argument. We used a convex model deliberately, and a higher capacity model would memorise more and would probably leak more, which would change the balance in favour of differential privacy. Finally, one site records no cholesterol at all and another has very few negative cases, so per site estimates at those two sites are less stable than the others.

4. CONCLUSION

Across four hospitals that genuinely never pooled their records, federated averaging recovered 87 percent of the benefit of pooling. That benefit was almost entirely confined to the site with the most missing data, which gained 0.050 in discrimination while no other site gained more than 0.001. A proximal correction for heterogeneity made no difference. Differential privacy degraded calibration several times faster than it degraded ranking, and cost between 1.6 and 2.7 times more when the model was applied to a hospital outside the training group. A membership inference attack recovered little from any model, and federated training leaked less than local training. Teams building clinical federations should report per site outcomes and calibration, not averages and discrimination alone.

ACKNOWLEDGEMENTS

The authors thank the investigators who collected and released the four hospital cardiology cohort, without which a study of a real rather than a simulated federation would not be possible.

This research received no specific grant from any funding agency in the public, commercial or not for profit sectors.

AUTHOR CONTRIBUTIONS

Abdul Ahad: Conceptualization, Methodology, Writing – Original Draft.

Mohammed Ali Shaik: Methodology, Supervision, Writing – Review & Editing.

Md Riyazuddin: Software, Data Curation, Writing – Review & Editing.

N Udaya Bhaskara Varma: Validation, Investigation, Writing – Review & Editing.

Md Jaffar Sadiq: Resources, Formal Analysis, Visualization.

All authors read and approved the final manuscript.

CONFLICT OF INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ETHICS STATEMENT

This study analysed a fully anonymised public research dataset that contains no identifiable patient information and is distributed for unrestricted research use. No new data were collected from patients. Ethical approval and individual consent were therefore not required for this secondary analysis.

Not applicable. The study uses a public de-identified dataset and reports no individual patient data, images or other personal details.

No generative artificial intelligence tool was used to produce the scientific content, the analysis or the results reported here.

DATA AVAILABILITY STATEMENT

The cardiology cohort analysed here is publicly available from the UCI Machine Learning Repository (Janosi and Steinbrunn, 1989). All analysis code and every result file needed to reproduce the tables and the figures are available from the corresponding author on reasonable request.

REFERENCES

- [1] Abadi, M., Chu, A., Goodfellow, I., et al., 2016. *Deep learning with differential privacy*, In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS). ACM, pp. 308–318. <https://doi.org/10.1145/2976749.2978318>
- [2] Bagdasaryan, E., Poursaeed, O., Shmatikov, V., 2019. *Differential privacy has disparate impact on model accuracy*, In: Advances in Neural Information Processing Systems (NeurIPS) 32, pp. 15479–15488.
- [3] Brisimi, T.S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I.C., Shi, W., 2018. *Federated learning of predictive models from federated electronic health records*. International Journal of Medical Informatics 112, 59–67. <https://doi.org/10.1016/j.ijmedinf.2018.01.007>
- [4] Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramèr, F., 2022. *Membership inference attacks from first principles*, In: Proceedings of the IEEE Symposium on Security and Privacy (S&P). IEEE, pp. 1897–1914. <https://doi.org/10.1109/SP46214.2022.9833649>
- [5] Collins, G.S., Reitsma, J.B., Altman, D.G., Moons, K.G.M., 2015. *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement*. BMJ 350, g7594. <https://doi.org/10.1136/bmj.g7594>
- [6] Dayan, I., Roth, H.R., Zhong, A., et al., 2021. *Federated learning for predicting clinical outcomes in patients with COVID-19*. Nature Medicine 27, 1735–1743. <https://doi.org/10.1038/s41591-021-01506-3>
- [7] Detrano, R., Janosi, A., Steinbrunn, W., et al., 1989. *International application of a new probability algorithm for the diagnosis of coronary artery disease*. American Journal of Cardiology 64(5), 304–310. [https://doi.org/10.1016/0002-9149\(89\)90524-9](https://doi.org/10.1016/0002-9149(89)90524-9)
- [8] Dwork, C., Roth, A., 2014. *The algorithmic foundations of differential privacy*. Foundations and Trends in Theoretical Computer Science 9(3–4), 211–407. <https://doi.org/10.1561/04000000042>
- [9] Geyer, R.C., Klein, T., Nabi, M., 2017. *Differentially private federated learning: A client level perspective*. arXiv preprint arXiv:1712.07557. <https://doi.org/10.48550/arXiv.1712.07557>
- [10] Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q., 2017. *On calibration of modern neural networks*, In: Proceedings of the 34th International Conference on Machine Learning (ICML). PMLR 70, pp. 1321–1330.
- [11] Holm, S., 1979. *A simple sequentially rejective multiple test procedure*. Scandinavian Journal of Statistics 6(2), 65–70.
- [12] Janosi, A., Steinbrunn, W., 1989. *Heart Disease [Dataset]*. UCI Machine Learning Repository. <https://doi.org/10.24432/C52P4X>
- [13] Kairouz, P., McMahan, H.B., Avent, B., et al., 2021. *Advances and open problems in federated learning*. Foundations and Trends in Machine Learning 14(1–2), 1–210. <https://doi.org/10.1561/22000000083>
- [14] Kaissis, G.A., Makowski, M.R., Rückert, D., Braren, R.F., 2020. *Secure, privacy-preserving and federated machine learning in medical imaging*. Nature Machine Intelligence 2(6), 305–311. <https://doi.org/10.1038/s42256-020-0186-1>
- [15] Li, T., Sahu, A.K., Talwalkar, A., Smith, V., 2020a. *Federated learning: Challenges, methods, and future directions*. IEEE Signal Processing Magazine 37(3), 50–60. <https://doi.org/10.1109/MSP.2020.2975749>
- [16] Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V., 2020b. *Federated optimization in heterogeneous networks*, In: Proceedings of Machine Learning and Systems (MLSys) 2, pp. 429–450.
- [17] McMahan, H.B., Moore, E., Ramage, D., Hampson, S., Agüera y Arcas, B., 2017. *Communication-efficient learning of deep networks from decentralized data*, In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). PMLR 54, pp. 1273–1282.
- [18] Mironov, I., 2017. *Rényi differential privacy*, In: Proceedings of the 30th IEEE Computer Security Foundations Symposium (CSF). IEEE, pp. 263–275. <https://doi.org/10.1109/CSF.2017.11>
- [19] Naeini, M.P., Cooper, G.F., Hauskrecht, M., 2015. *Obtaining well calibrated probabilities using Bayesian binning*, In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. AAAI Press, pp. 2901–2907.
- [20] Nasr, M., Shokri, R., Houmansadr, A., 2019. *Comprehensive privacy analysis of deep learning*, In: Proceedings of the IEEE Symposium on Security and Privacy (S&P). IEEE, pp. 739–753. <https://doi.org/10.1109/SP.2019.00065>

- [21] Niculescu-Mizil, A., Caruana, R., 2005. *Predicting good probabilities with supervised learning*, In: Proceedings of the 22nd International Conference on Machine Learning (ICML). ACM, pp. 625–632. <https://doi.org/10.1145/1102351.1102430>
- [22] Pedregosa, F., Varoquaux, G., Gramfort, A., et al., 2011. *Scikit-learn: Machine learning in Python*. Journal of Machine Learning Research 12, 2825–2830.
- [23] Rieke, N., Hancox, J., Li, W., et al., 2020. *The future of digital health with federated learning*. npj Digital Medicine 3, 119. <https://doi.org/10.1038/s41746-020-00323-1>
- [24] Riley, R.D., Ensor, J., Snell, K.I.E., et al., 2020. *Calculating the sample size required for developing a clinical prediction model*. BMJ 368, m441. <https://doi.org/10.1136/bmj.m441>
- [25] Sheller, M.J., Edwards, B., Reina, G.A., et al., 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>
- [26] Shokri, R., Stronati, M., Song, C., Shmatikov, V., 2017. *Membership inference attacks against machine learning models*, In: Proceedings of the IEEE Symposium on Security and Privacy (S&P). IEEE, pp. 3–18. <https://doi.org/10.1109/SP.2017.41>
- [27] Steyerberg, E.W., Vickers, A.J., Cook, N.R., et al., 2010. *Assessing the performance of prediction models: A framework for traditional and novel measures*. Epidemiology 21(1), 128–138. <https://doi.org/10.1097/EDE.0b013e3181c30fb2>
- [28] Truex, S., Baracaldo, N., Anwar, A., et al., 2019. *A hybrid approach to privacy-preserving federated learning*, In: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (AISec). ACM, pp. 1–11. <https://doi.org/10.1145/3338501.3357370>
- [29] Vaid, A., Jaladanki, S.K., Xu, J., et al., 2021. *Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: Machine learning approach*. JMIR Medical Informatics 9(1), e24207. <https://doi.org/10.2196/24207>
- [30] Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W., 2019. *Calibration: The Achilles heel of predictive analytics*. BMC Medicine 17, 230. <https://doi.org/10.1186/s12916-019-1466-7>
- [31] Warnat-Herresthal, S., Schultze, H., Shastry, K.L., et al., 2021. *Swarm learning for decentralized and confidential clinical machine learning*. Nature 594, 265–270. <https://doi.org/10.1038/s41586-021-03583-3>
- [32] Xu, J., Glicksberg, B.S., Su, C., Walker, P., Bian, J., Wang, F., 2021. *Federated learning for healthcare informatics*. Journal of Healthcare Informatics Research 5(1), 1–19. <https://doi.org/10.1007/s41666-020-00082-4>
- [33] Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S., 2018. *Privacy risk in machine learning: Analyzing the connection to overfitting*, In: Proceedings of the 31st IEEE Computer Security Foundations Symposium (CSF). IEEE, pp. 268–282. <https://doi.org/10.1109/CSF.2018.00027>
- [34] Zhao, Y., Li, M., Lai, L., Suda, N., Civin, D., Chandra, V., 2018. *Federated learning with non-IID data*. arXiv preprint arXiv:1806.00582. <https://doi.org/10.48550/arXiv.1806.00582>