2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3315-8618-8/25/$31.00 ©2025 Authors | DOI: 10.1109/EMBC58623.2025.11252903

Hospital Participation in Federated Learning: Evaluating Sustainability and Clinical Utility

Andrei Kazlouski
Department of Computing
University of Turku
Turku, Finland
ankazl@utu.fi

Ileana Montoya Perez
Department of Computing
University of Turku
Turku, Finland
iimope@utu.fi

Tapio Pahikkala
Department of Computing
University of Turku
Turku, Finland
aatapa@utu.fi

Antti Airola
Department of Computing
University of Turku
Turku, Finland
ajairo@utu.fi

***Abstract*—Prostate cancer (PCa) diagnosis often relies on biopsies, which can lead to unnecessary procedures and complications. Federated learning (FL) offers a privacy-preserving approach for training predictive models across hospitals without sharing sensitive patient data. In this study, we evaluate the feasibility of FL for PCa risk prediction by benchmarking different training strategies, including local, federated models, as well as free-riding (FR) on federated models. Using real-world heterogeneous datasets from 19 hospitals, we analyze the impact of data diversity and consortium size on predictive performance. Our results show that while FL improves model generalizability, local models often perform comparably, making direct participation in FL less beneficial for large hospitals. However, a small consortium of high-data-quality institutions could collaboratively develop robust models for broader clinical use. We discuss the practical implications of FL in healthcare and propose strategies for sustainable deployment in real-world hospital networks.**

***Index Terms*—Federated Learning, Prostate Cancer, Health informatics, Privacy-preserving, Open Source**

I. Introduction

Federated Learning (FL) is gaining recognition among clinicians as a way to integrate diagnostic insights from hospitals worldwide without sharing individual patient data. Studies show that FL can achieve results comparable to centralized learning, where data from all hospitals are combined. However, real-world adoption remains limited due to setup complexity and regulatory challenges.

For example, Finland's secondary use act[1] strictly regulates how patient data can be used beyond diagnosis. As a result, it can take years to establish a functioning federated consortium. To justify participation, hospitals must see clear benefits for diagnosing their local population compared to alternative methods. Hospitals generally have three options for obtaining diagnostic models: training one locally, jointly training with others using FL, or using an externally trained model (see Figure 1 for an overview).

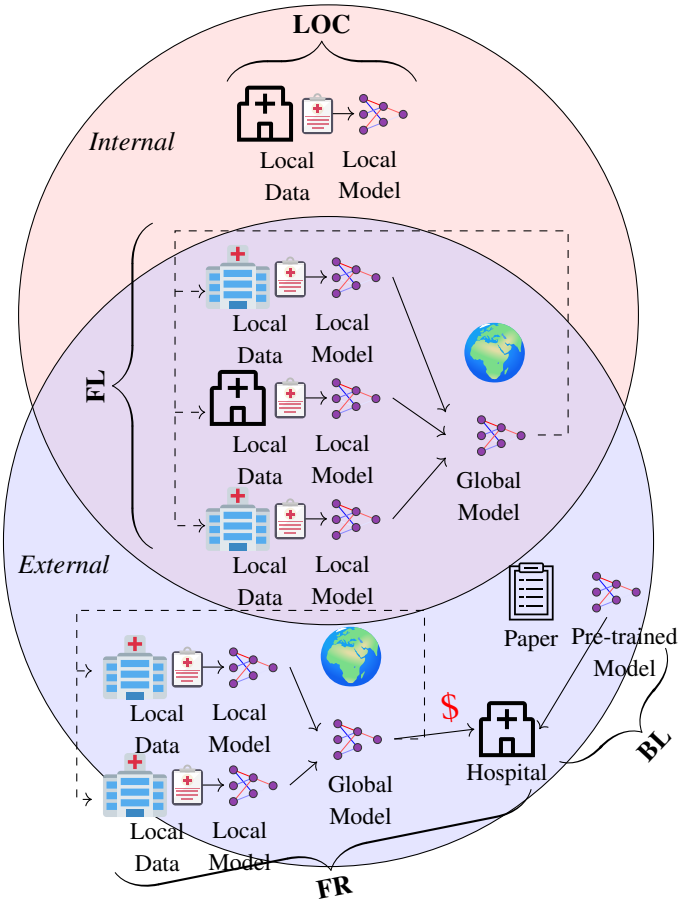

Fig. 1: A hospital can obtain a predictive model by four distinct strategies: Local Training (LOC) — training a model solely on the hospital's own data; Federated Learning (FL) — contributing data to a collaborative, distributed training process; Free-Riding (FR) — using a 'pre-trained' federated model without contributing data; and Baseline (BL) — relying on an external pre-trained model developed independently of the hospital's data.

This work has received funding from European Union's Horizon Europe research and innovation programme (grant number 101095384) and from Research Council of Finland (grants 358868, 345804, 345805, 340140, 340182). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them.

[1]https://findata.fi/en/services-and-instructions/legislation/#What-is-the-Secondary-Use-Act

Very small hospitals with insufficient patient data contribute little to the global model, making their participation impractical. A more sustainable approach could involve larger hospitals with high-quality data training and maintaining a federated model for others to use. In literature, this practice is called "free-riding" (FR), often viewed negatively as an unfair exploitation of others' efforts.

However, if properly structured and compensated, FR could enable privacy-preserving health data sharing. This study examines the sustainability of such setups by exploring different training strategies and analyzing how the number of participating hospitals affects FL model performance. Simply put, we investigate how many hospitals are needed for an FL model to be effective in external medical institutions.

Most FL studies in healthcare rely on artificially partitioned datasets or challenges, which often lack sufficient documentation. While useful for preliminary validation, these approaches do not fully capture the heterogeneity and imbalances of real-world hospital data. Some studies have distributed FL training across hospitals or used geographically diverse medical data, but the number of federated clients rarely exceeds 10 – making it difficult to analyze learning curves based on hospital count.

Notably, Patti et al. [13] and Kerkouche et al. [3] included 71 and 415 sites, respectively, but did not examine learning curves in this context. To address these gaps and better represent federated client diversity, we used 19 public datasets from nine countries across Europe, America, and Asia. We employ the same prostate cancer data collection, as used in [2]. These datasets come from peer-reviewed medical publications that detail medical tasks, procedures, and data processing methods.

We apply our methodology to predicting prostate cancer (PCa) biopsy outcomes. Prostate biopsies are crucial for diagnosis but can lead to serious complications such as bleeding and infection. Since many biopsies reveal benign or low-risk cancer, reducing unnecessary procedures is a major goal in modern urology.

Several models use MRI and clinical variables to predict biopsy outcomes [1, 11, 15, 14]. While these models have demonstrated high sensitivity and can accurately predict the presence of high-risk PCa in biopsy results, their specificity remains suboptimal, as some significant PCa cases are still being missed. Therefore, improving and comparing models to enhance predictive accuracy is crucial.

While FL between hospitals generally suffers from data heterogeneity, this challenge is especially pronounced in prostate cancer due to differences in MRI equipment across institutions and variability introduced by human radiologists when assigning scores using the Prostate Imaging Reporting and Data System (PI-RADS). These factors make FL particularly difficult in this domain.

To summarize, the key contributions of our work are as follows:

- **Sustainable FL framework**: We propose and evaluate a realistic approach for setting up sustainable FL consortia in cross-hospital settings. We investigate practical strategies to help institutions develop high-performing diagnostic models tailored to their patient populations.
- **Reproducible benchmarks:** Our benchmarks are fully reproducible, with all data and code publicly available[2].

II. Methods

To evaluate the sustainability of FL setups for external medical institutions in PCa research, we used two risk calculators. Their architectures and provided coefficients served as baseline pre-trained models.

Since this is a simulated study, all federated training was performed on a single machine without actual data distribution. However, the setup replicates real-world FL

[2]https://github.com/phaseivai/Hospital-Participation-in-FL

TABLE I: Public datasets and their corresponding articles are listed. The studied datasets are referred to by their alpha-2 country code and order number (e.g., es2). The features are defined as follows: PSA = prostate-specific antigen, PSAD = PSA density, PV = prostate volume, fPSA = free PSA, DRE = digital rectal examination, PB = previous biopsy, FH = family history, PI-RADS = Prostate Imaging Reporting and Data System.

Country	Dataset	Study	Patient #	Age	PSA	PV	PSAD	fPSA	PB	DRE	FH	Race	PI-RADS	ISUP
Spain	es1	[9]	79	✓	✓	✓	✓	✗	✓	✓	✓	✗	2.0	✓
	es2		16										2.0	✓
	es3		30										2.0	✓
	es4		200										2.0	✓
	es5		200										2.0	✓
	es6		346										2.0	✓
	es7		711										2.0	✓
	es8		58										2.0	✓
Korea	kr1	[5]	406	✓	✓	✓	✓	✗	✓	✗	✗	✗	2.0	✓
	kr2	[10]	590	✓	✓	✓	✓	✗	✗	✓	✗	✗	2.1	✓
Germany	de1	[4]	763	✓	✓	✓	✓	✗	✗	✗	✗	✗	2.1	✓
	de2	[15]	162	✓	✓	✓	✓	✗	✗	✓	✗	✗	2.0	✓
China	ch2	[12]	530	✓	✓	✓	✓	✓	✗	✗	✗	✗	2.0	✓
	ch1	[17]	312	✓	✓	✓	✓	✓	✓	✓	✗	✗	1.0	✓
USA	us2	[16]	310	✓	✓	✓	✓	✓	✓	✓	✓	✓	2.0	✓
	us1	[8]	280	✓	✓	✓	✓	✗	✗	✓	✓	✓	1.0/2.0/2.1	✓
Netherlands	nl	[17]	266	✓	✓	✓	✓	✓	✓	✓	✗	✗	1.0	✓
Italy	it	[6]	218	✓	✓	✓	✓	✗	✗	✗	✗	✗	2.0	✓
UK	gb	[15]	133	✓	✓	✓	✓	✗	✗	✓	✗	✗	2.0	✓

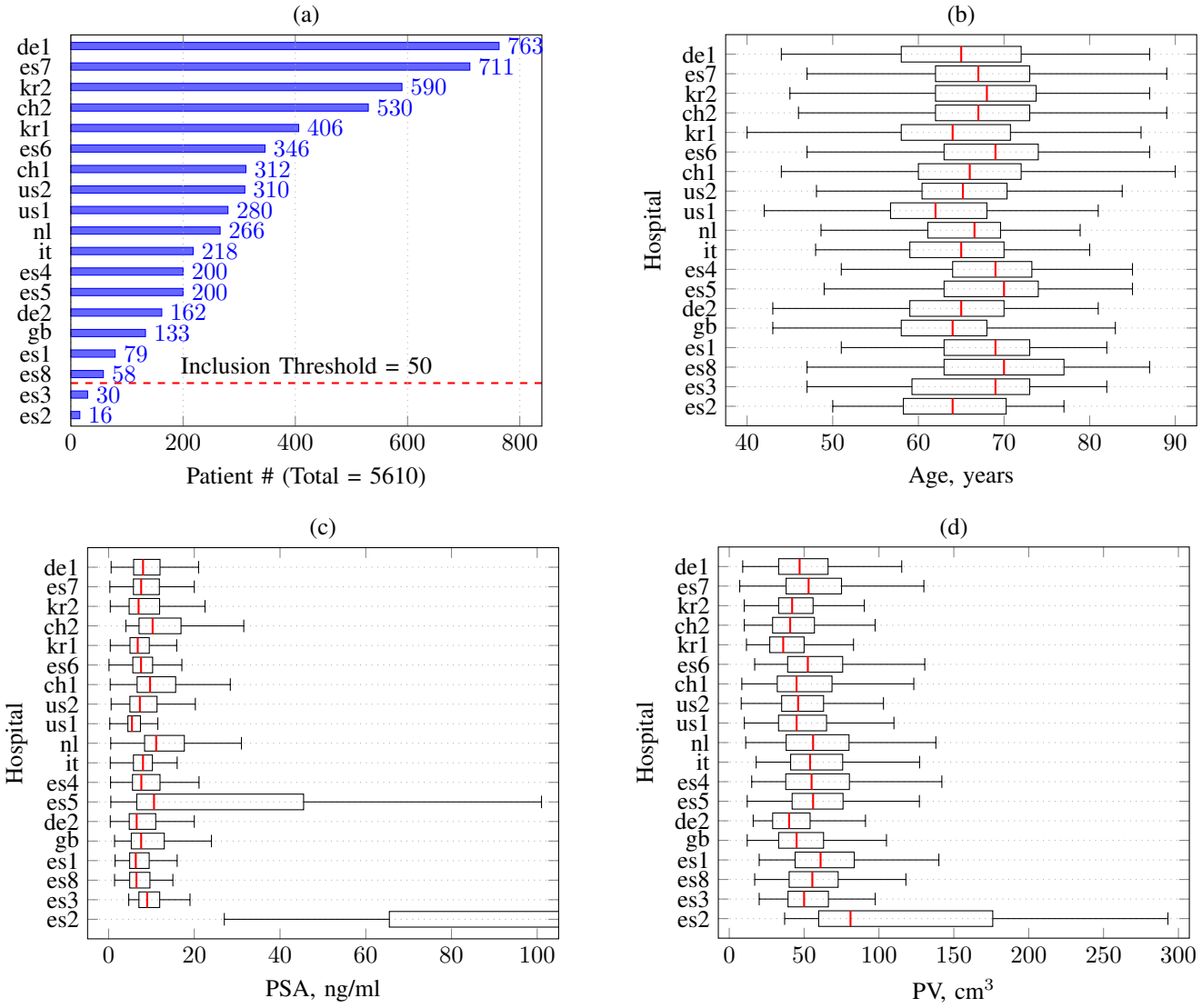

Fig. 2: Statistics of numerical parameters across the studied datasets.

scenarios, where institutions collaboratively train models without sharing raw patient data. We used the Federated Averaging (FedAvg) [7] algorithm for model aggregation, a widely adopted approach in FL due to its efficiency and scalability.

We compared models trained using FL and FR with locally trained and externally sourced models to assess their clinical utility. Additionally, we conducted experiments under different federation conditions, analyzing how the number of participating hospitals affects predictive accuracy. For each experiment repetition, all modeling decisions were made based on the *training* sets. This research provides insights into the feasibility of FL for PCa diagnostics and suggests strategies for its sustainable implementation in real-world hospital networks.

A. Data

Table II summarizes the clinical parameters for each dataset, along with their country of origin and corresponding publication. Typically, hospitals with fewer than 50 patients would not qualify to participate in federated training. However, since the models used in this study are simple logistic regressions (see Section II-C), one could argue that including them is still reasonable. Moreover, because one of the main goals of this paper is to study the learning curves of federated learning, we chose to *retain* the smaller hospitals in our analysis.

To maintain consistency, all datasets were standardized to use the same units for key parameters. Samples with missing values were discarded, and appropriate preprocessing and harmonization steps were applied. Most variables are unambiguous, and it was verified that they were measured using consistent units across sites. The most challenging variable to standardize is PI-RADS, as different system versions and the subjectivity of human radiologist scorers can lead to inconsistencies, as shown in Table I.

Figure 2 provides a breakdown of the key numerical features common across all datasets, sorted from the largest to

TABLE II: Risk calculators that include parameters present in the studied datasets, along with the papers that introduce them.

Used	Calculator	Study	Histopathology	Age	PSA	PV	PI-RADS	PSAD	fPSA	PB	DRE	FH	Race
✓	Ettala	[1]	ISUP ≥ 3	✓	✓	✓	✓						
✓	Noh	[11]	ISUP ≥ 2	✓			✓	✓					
✗	Radtke	[15]	ISUP ≥ 2	✓	✓	✓	✓				✓		
✗	Peters	[14]	ISUP ≥ 3	✓	✓	✓	✓			✓	✓	✓	✓

Train \ Test	de1	es7	kr2	ch2	kr1	es6	ch1	us2	us1	nl	it	es4	es5	de2	gb	es1	es8	es3	es2	AV	AVW
BL	83.5	87.7	83.2	88.3	86.5	78.0	85.1	75.4	70.3	93.5	83.5	89.5	71.9	78.3	90.4	50.7	78.4	71.3	58.3	79.2	83.0
LOC	85.5	89.2	87.6	89.8	90.5	80.5	85.3	79.7	78.5	91.2	76.6	89.9	77.2	78.7	87.8	38.6	74.7	69.3	72.3	80.1	84.7
CEN	85.3	89.1	87.3	89.0	90.0	80.8	85.2	79.7	77.1	91.3	82.4	91.8	77.6	80.3	91.2	49.9	81.3	79.8	60.5	81.6	85.2
FL	84.3	87.5	87.7	87.2	90.7	81.0	84.6	80.8	78.4	87.5	81.0	91.6	69.8	79.6	90.2	50.6	80.7	82.3	53.9	80.5	84.3
FR	83.8	86.7	87.7	86.2	90.6	80.8	84.5	80.8	78.2	87.3	80.8	91.6	69.8	79.3	90.2	51.2	80.3	81.6	54.1	80.3	84.0

(a) Ettala

Train \ Test	de1	es7	kr2	ch2	kr1	es6	ch1	us2	us1	nl	it	es4	es5	de2	gb	es1	es8	es3	es2	AV	AVW
BL	83.6	88.2	87.2	88.9	89.8	80.9	85.5	77.7	77.1	93.4	82.5	90.5	72.4	81.6	90.7	55.2	82.3	74.5	90.0	82.7	84.7
LOC	84.9	88.0	87.0	89.0	89.4	81.1	86.3	76.7	77.1	93.1	82.2	89.8	75.6	81.0	90.7	48.3	79.2	65.7	90.9	81.9	84.7
CEN	84.2	88.2	87.2	89.0	89.8	80.7	85.2	77.8	76.9	93.4	82.8	90.4	73.5	81.7	91.4	54.5	81.8	75.2	89.6	82.8	84.8
FL	83.8	88.2	87.2	88.9	89.8	80.9	85.4	77.8	77.1	93.4	82.6	90.5	72.9	81.7	91.1	55.1	81.9	74.9	89.6	82.8	84.8
FR	83.3	88.1	87.2	88.9	89.9	80.7	85.2	77.8	77.0	93.4	82.4	90.5	72.7	81.7	91.0	54.9	81.9	74.8	89.6	82.7	84.7

(b) Noh

Max (100)

AUC

Min (40)

Fig. 3: Training strategies for the Ettala (a) and Noh (b) calculators. Rows represent different training strategies: BL – pre-trained baseline, LOC – locally trained model, CEN – centrally trained model (using all available data), FL – federated training, and FR – free-riding a federated model, tested exclusively on one excluded dataset. Columns indicate the datasets used for testing. AV represents the average performance per hospital, while AVW is the per-patient weighted equivalent. Datasets are sorted from the largest to the smallest.

the smallest hospital. While some differences exist, the most notable outlier is the Spain2 dataset, where both PSA and PV values significantly exceed the global average. For this study clinically significant prostate cancer (csPCa) is defined as a histopathological International Society of Urological Pathology (ISUP) Grade Group ≥ 2 observed in prostate biopsies.

B. Risk Calculators

To ensure compatibility across datasets, we selected risk calculators that include variables present in at least one dataset. Table II lists four such models, but only the first two – Ettala and Noh [1, 11] – contain variables common to all datasets.

Therefore, we used the architectures of Ettala and Noh as the foundation for our study. These models serve as representative risk calculators for predicting PCa biopsy outcomes in a federated setting. Notably, the original Ettala baseline model uses different cutoffs for csPCa, addressing slightly different clinical questions.

C. Models

Both baseline models use logistic regression for binary classification of benign versus csPCa. Both models rely on the base variables or their transformations listed in Table II.

III. Results

We begin our experiments by training both baseline model architectures using all training strategies for each hospital to generate a prediction model. The results are shown in Figure 3 as two heatmap matrices, illustrating AUCs for both model architectures. The rows represent the four training strategies – Baseline (BL), Local (LOC), FL, and Free-Riding (FR) – along with Centralized Training (CEN) as an optimal reference point. The columns correspond to 19 distinct test sets, along with their row-wise average (AV) and its patient-weighted equivalent (AVW).

For the FR strategy, all datasets except the one used for testing were included in training. The results are averaged over 10,000 Monte Carlo repetitions. As before, all modeling decisions were based on the training sets – 80% of the data for all training strategies, except for FR and BL, where the entire dataset was used for evaluation.

Overall, it appears that CEN outperforms all other training strategies (as expected). The baseline model for Ettala performs noticeably worse than other strategies, whereas for Noh, it is almost on par with centralized training. The weaker performance of the Ettala baseline model can likely be attributed to differences in the clinical problem it was originally trained to address. FL slightly outperforms FR, which is expected since both models are similar, but FL benefits from an additional participating hospital. Interestingly, local models are, on average, only slightly worse than FL or FR. In larger hospitals, local models often outperform other training strategies, reinforcing the idea that a consortium of a few large hospitals could effectively collaborate to train and maintain an FL model for broader institutional use.

Comparing the Ettala and Noh calculators, Noh consistently outperforms Ettala across all training strategies.

The most notable difference between the two models is observed in the Spain2 (es2) dataset. While Noh correctly identifies negative cases, Ettala fails to do so. A closer look at the data reveals extreme PSA outliers in this dataset (Figure 2). Specifically, 9 out of 16 patients have PSA levels above 100 ng/mL, with values reaching as high as 1891 ng/mL, yet only 3 of these patients have csPCa.

Since the Ettala calculator applies a spline transformation, it heavily penalizes these extreme PSA values, leading to overprediction of csPCa. In practice, however, patients with such exceptionally high PSA levels would likely be recommended for a biopsy regardless of the calculator's output.

Another notable case is the Spain1 (es1) dataset, where both calculators fail to predict outcomes accurately for the local population. Even the locally trained models, which are limited by a small sample size of only 79 patients, struggle to distinguish between labels. This likely indicates irregularities or some other unmeasured factors.

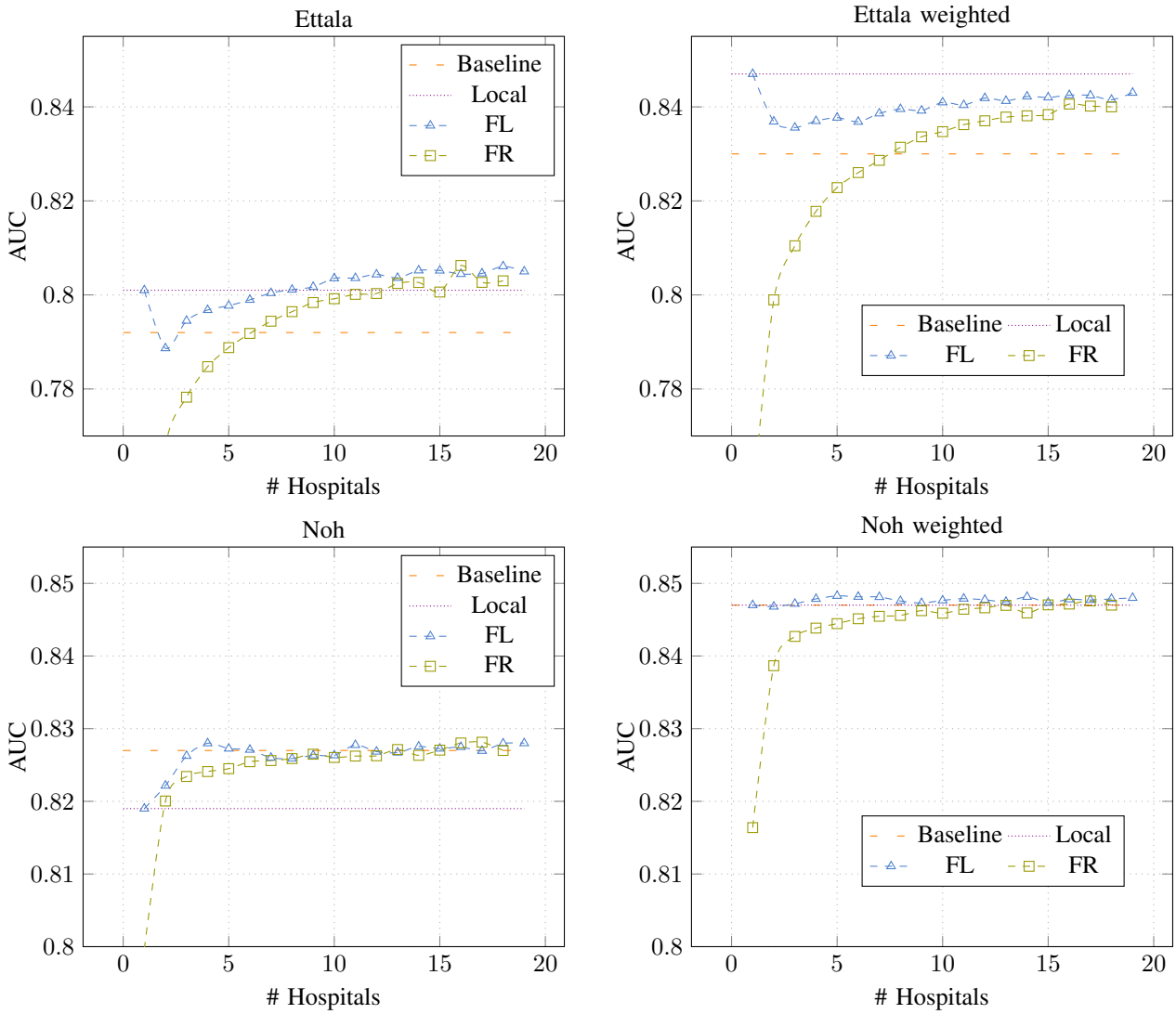

Fig. 4: Learning curves for different training methods based on the number of datasets and their per-patient weighted equivalents. The Baseline and Local curves remain flat since these models do not involve federation. FL curves show the average performance across participating hospitals, while FR curves represent performance on excluded hospitals.

Overall, a detailed clinical analysis of these calculators is left for future work.

We now turn to the paper's main research question by analyzing the learning curves of different training strategies.

To assess the feasibility of establishing a stable FL setup for broader hospital use, we examine the learning curve of FR based on the total number of participating hospitals. For comparison, we also plot learning curves for other training methods. These results are shown in Figure 4.

Since the Baseline and Local models do not depend on the number of hospitals, their curves remain horizontal and are taken directly from Figure 3. The FL and FR curves, however, are computed using 30,000 Monte Carlo simulations. For each simulation, we randomly select a number of datasets (ranging from 1 to 19) with equal probability. We then train an FL model and evaluate it on holdout sets from the participating datasets for the FL curve, while for the FR curve, we test on the excluded datasets. The final curves represent the averaged results across all simulations (see Algorithm 1 for an overview).

For certain FL and FR values, simulations were not performed, as the results can be inferred from Figure 3. Specifically, FL(1) corresponds to local training, FL(19) is directly displayed in the table, FR(0) is set to 0.5 (as no hospitals contribute to training), and FR(18) is also shown in the same figure. FR(19) does not exist, as there are no free-riders remaining.

Since the difference between FL(1) and FL(19) is relatively small, the FL curve shows only a slight and inconsistent improvement, making participation in FL less appealing for hospitals. In contrast, the AUC for FR steadily improves as the number of participating hospitals increases. Once around 10 hospitals are involved, the FR model consistently performs on par with other training approaches.

It is important to note that hospitals were selected randomly, with no preference given to larger institutions. If larger hospitals were prioritized, fewer participants would likely be needed to achieve similar average AUCs. This suggests that collaborative training could produce models that generalize well to other hospitals.

If these collaboratively maintained models outperform locally trained and pre-trained baseline models (as our results suggest), they could present a viable business opportunity. Smaller hospitals might be willing to pay for access to such models rather than investing in their own training. Given the trade-offs between setup complexity, privacy concerns, and potential benefits, direct participation in FL may not be the optimal choice for many hospitals. Since comparable models could be obtained with a small financial investment, the added complexity of FL may not be justified.

Algorithm 1 Hospital Participation in FL and FR

Input: Datasets $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, number of Monte Carlo simulations $num\_repeats$
Output: Average FL AUCs and FR AUCs for each dataset count $r \in \{1, \ldots, N\}$

for repeat $\leftarrow 1$ to $num\_repeats$ **do**
  $dataset\_count \leftarrow$ randomly sampled count $r \in \{1, \ldots, N\}$
  Randomly shuffle $\mathcal{D}$
  Select the first $dataset\_count$ datasets as $participating\_datasets$
  Let $excluded\_datasets$ be the remaining datasets in $\mathcal{D}$
  Simulate federated training and evaluation
  Initialize server and clients for federated training
  for each round **do**
    for each client in $participating\_datasets$ **do**
      Train models on 80% train split (random)
      Collect model updates from clients
      Aggregate client updates and broadcast to all clients
    end for
  end for
  Evaluate AUC on 20% test set for each dataset in $participating\_datasets$
  Evaluate AUC on 100% of data for each dataset in $excluded\_datasets$
end for
Compute mean and weighted mean of AUCs for $dataset\_count$
Compute mean and weighted mean of AUCs for $excluded\_count$
return AUCs for FL and FR

IV. Discussion and Conclusion

A key takeaway from our results is that, in real-world settings, participation in FL offers diminishing returns for hospitals, as local models perform nearly as well. If enough contributors are already part of a federated learning consortium, further participation in training the global model becomes less practical. However, if every hospital opts out of federated training, there will be no model available to share with others. As emphasized throughout this work, a consortium of trusted maintainers with high-quality and sufficient data is crucial for providing robust models to other hospitals.

Realism of the setup. The practicality of the described setup may seem overly specific, but it is widely considered in practice. For instance, the European Health Data Space (EHDS), a health-specific ecosystem that includes rules, common standards, practices, infrastructure, and a governance framework[3], identifies providing a trustworthy and efficient setup for using health data as one of its main goals. Federated learning and medical models as a service are central components of this system. Our proposed approach to practical free-riding scenarios aligns with the EHDS framework's objectives for the secondary use of health data, where 'pre-trained' federated learning models may be made available to *other* institutions for commercial or research purposes.

Limitation of the Study. While methodological differences,

[3]https://www.european-health-data-space.com/

patient variability, and highly inconsistent population sizes across the studied datasets could be seen as limitations, we consider them a strength. As noted earlier, many FL studies overlook real-world heterogeneity by artificially splitting data among clients. Instead of simplifying our approach by smoothing out these differences, we embrace them to better simulate the diversity of hospitals that might use the final federated model. This setup allows us to explore the most effective training strategies under real-world conditions.

A genuine limitation of our study is the relatively small number of participating hospitals. Although this number is comparable to or even larger than those in other FL studies in healthcare, having more hospitals would allow us to construct even more accurate learning curves. Additionally, we do not explore the impact of the number of participating sites on model bias or the potential deterioration of fairness in FL models. As this is a fundamental challenge in any federated learning application, we leave it as an open direction for future research.

Overall, our findings suggest that a collaborative setup, where well-established hospitals jointly train models and share them with other institutions, holds significant promise. Such models can achieve higher accuracy than those available to smaller institutions individually. However, the optimal training strategy for each hospital is likely context-dependent, and alternative approaches should also be explored.

References

[1] Otto Ettala et al. "Individualised non-contrast MRI-based risk estimation and shared decision-making in men with a suspicion of prostate cancer: Protocol for multicentre randomised controlled trial (multi-IMPROD V. 2.0)". In: *BMJ open* 12.4 (2022), e053118.

[2] Andrei Kazlouski et al. "Towards Practical Federated Learning and Evaluation for Medicalprediction Models". In: *Available at SSRN 5119417* (2025).

[3] Raouf Kerkouche et al. "Privacy-preserving and bandwidth-efficient federated learning: An application to in-hospital mortality prediction". In: *Proceedings of the conference on health, inference, and learning*. 2021, pp. 25–35.

[4] M Klingebiel et al. "Data on the detection of clinically significant prostate cancer by magnetic resonance imaging (MRI)-guided targeted and systematic biopsy". In: *Data in Brief* 45 (2022), p. 108683.

[5] Seung Soo Lee et al. "Usefulness of Bi-Parametric Magnetic Resonance Imaging with b= 1,800 s/mm2 Diffusion-Weighted Imaging for Diagnosing Clinically Significant Prostate Cancer". In: *The World Journal of Men's Health* 38.3 (2020), p. 370.

[6] Eugenio Martorana et al. "Lesion volume predicts prostate cancer risk and aggressiveness: validation of its value alone and matched with prostate imaging reporting and data system score". In: *BJU international* 120.1 (2017), pp. 92–103.

[7] Brendan McMahan et al. "Communication-efficient learning of deep networks from decentralized data". In: *Artificial intelligence and statistics*. PMLR. 2017, pp. 1273–1282.

[8] Julio Meza et al. "Assessing the accuracy of multiparametric MRI to predict clinically significant prostate cancer in biopsy naïve men across racial/ethnic groups". In: *BMC urology* 22.1 (2022), p. 107.

[9] Juan Morote et al. "The Role of Digital Rectal Examination Prostate Volume Category in the Early Detection of Prostate Cancer: Its Correlation with the Magnetic Resonance Imaging Prostate Volume". In: *The World Journal of Men's Health* 42.2 (2024), p. 441.

[10] Jong Kil Nam et al. "Impact of Ultrasonographic Findings on Cancer Detection Rate during Magnetic Resonance Image/Ultrasonography Fusion-Targeted Prostate Biopsy". In: *The World Journal of Men's Health* 41.3 (2023), p. 743.

[11] Tae Il Noh et al. "A predictive model based on biparametric magnetic resonance imaging and clinical parameters for clinically significant prostate cancer in the Korean population". In: *Cancer Research and Treatment: Official Journal of Korean Cancer Association* 53.4 (2021), pp. 1148–1155.

[12] Jin-feng Pan et al. "Modified predictive model and nomogram by incorporating prebiopsy biparametric magnetic resonance imaging with clinical indicators for prostate biopsy decision making". In: *Frontiers in Oncology* 11 (2021), p. 740868.

[13] Sarthak Pati et al. "Federated learning enables big data for rare cancer boundary detection". In: *Nature communications* 13.1 (2022), p. 7346.

[14] Max Peters et al. "Predicting the need for biopsy to detect clinically significant prostate cancer in patients with a magnetic resonance imaging–detected prostate imaging reporting and data system/Likert $\geq$ 3 lesion: development and multinational external validation of the imperial rapid access to prostate imaging and diagnosis risk score". In: *European urology* 82.5 (2022), pp. 559–568.

[15] Jan Philipp Radtke et al. "Prediction of significant prostate cancer in biopsy-naïve men: Validation of a novel risk model combining MRI and clinical parameters and comparison to an ERSPC risk calculator and PI-RADS". In: *PLoS One* 14.8 (2019), e0221350.

[16] Ardeshir R Rastinehad et al. "Comparison of multiparametric MRI scoring systems and the impact on cancer detection in patients undergoing MR US fusion guided prostate biopsies". In: *PloS one* 10.11 (2015), e0143404.

[17] Kai Zhang et al. "Distribution of Prostate Imaging Reporting and Data System score and diagnostic accuracy of magnetic resonance imaging–targeted biopsy: comparison of an Asian and European cohort". In: *Prostate International* 7.3 (2019), pp. 96–101.