International Journal of Medical Informatics 204 (2025) 106046

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Towards practical federated learning and evaluation for medical prediction models

Andrei Kazlouski [a,*], Ileana Montoya Perez [a], Faiza Noor [a], Mikael Högerman [b,c,d], Otto Ettala [b,c], Tapio Pahikkala [a], Antti Airola [a]

[a] Department of Computing, University of Turku, Turku, Finland
[b] Department of Urology, Turku University Hospital, Turku, Finland
[c] Department of Urology, University of Turku, Turku, Finland
[d] Department of Mathematics and Statistics, University of Turku, Turku, Finland

ARTICLE INFO

Dataset link: https://github.com/phaseivai/Towards-Practical-FL/

Keywords:
Cross-silo FL
Federated evaluation
Federated learning
Open access
Prostate cancer

ABSTRACT

Background: Federated learning (FL) is a rapidly advancing technique that enables collaborative model training while preserving data privacy. This approach is particularly relevant in healthcare, where privacy concerns and regulatory restrictions often prevent centralized data sharing. FL has shown promise in tasks such as disease detection, achieving performance levels comparable to centralized systems. However, its practical usability in real-world applications remains underexplored.

Methods: We evaluate the practical effectiveness of FL in predicting whether patients suspected of prostate cancer require invasive biopsy procedures. The study uses 14 publicly available prostate cancer datasets from 10 countries. We propose and benchmark a novel FL evaluation strategy, Leave-Silo-Out (LSO), which quantifies the performance gap between federated training and free-riding (utilizing the federated model without contributing data). Additionally, we investigate whether locally trained models can outperform multi-hospital FL models. The results are assessed with a focus on improving the diagnosis of local patients.

Results: Our findings reveal that the benefits of FL vary with the amount of locally available annotated data. Hospitals with very small datasets see negligible improvements from FL compared to free-riding. Institutions with moderate datasets may achieve some gains through FL training. However, hospitals with extensive datasets often experience little to no advantage from FL and, in some cases, observe reduced performance compared to local training.

Conclusion: Federated learning shows potential in scenarios with limited data availability. However, its practical applicability is highly context-dependent, influenced by factors such as data availability and specific task requirements.

1. Introduction

Federated Learning (FL) is gaining recognition as a secure method of training models across multiple data sources. Its core principle is that sites share only model parameters, not raw data. This enables collaborative model development while preserving the privacy of individual datasets. FL algorithms are categorized into cross-device FL for training on mobile/IoT devices and cross-silo FL for training with a few stable, trusted clients [8]. Cross-silo FL has gained particular interest in healthcare, enabling hospitals to collaborate on disease diagnosis without exposing patient information.

Although cross-device FL has been widely adopted in real-world applications, cross-silo FL, especially cross-hospital, remains rare outside of research settings. Regulatory challenges are a significant barrier to implementing cross-silo FL [24], even for cohorts within the same country, let alone internationally. Another challenge lies in demonstrating tangible benefits to individual hospitals participating in a multi-silo cohort.

Many cross-silo FL studies rely on artificially partitioned homogeneous benchmark datasets, such as CIFAR-10 or MNIST, divided among simulated clients. Although useful for preliminary validation, this approach fails to capture the real-world complexities and data heterogeneity of cross-silo FL applications, particularly in healthcare.

* Corresponding author.
E-mail address: ankazl@utu.fi (A. Kazlouski).

https://doi.org/10.1016/j.ijmedinf.2025.106046
Received 24 January 2025; Received in revised form 8 July 2025; Accepted 9 July 2025
Available online 18 July 2025

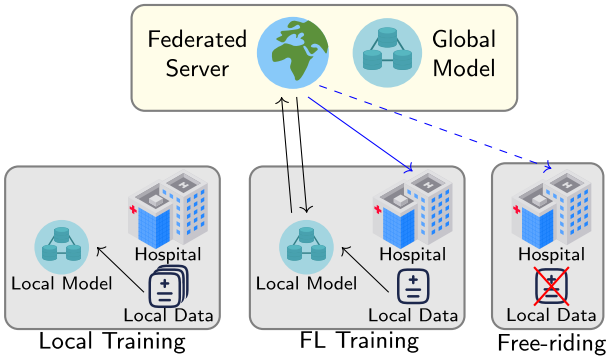


Fig. 1. Federated training strategies for real-world medical applications: Local, FL, and Free-riding (FR). Blue arrows indicate the final global model being shared to the hospitals. The preferred strategy depends on the amount of local data. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

When datasets originate from different hospitals, they often lack critical documentation on essential factors such as data collection methods, equipment, and procedural differences. Although public datasets for FL are sometimes accessible through platforms such as Kaggle competitions [38] or specialized medical challenges [37], comparing them can feel like "comparing apples and oranges" due to inconsistencies between sites.

Even when data is collected across multiple institutions with thorough documentation, evaluation methodologies often remain underdeveloped for FL-specific tasks. Although many studies show that FL models can achieve performance comparable to centralized training (where data from all sites is combined) [36], there is limited analysis on when FL benefits a hospital enough to justify participation.

Broadly, hospitals have three main options to obtain diagnostic models:

- Local Training – training exclusively on their own data.
- FL Training – actively participating in federated learning.
- Free-riding (FR) – using a federated model without contributing data.

In the literature, the term free-rider [12] often carries a negative connotation, implying the unfair exploitation of others' efforts. However, for smaller institutions, relying on shared predictive models may be their only option, as they lack sufficient data to train their own and contribute negligible amounts to the global model.

In this paper, we propose a comprehensive evaluation methodology to assess the practical usability of FL. We introduce a novel evaluation strategy, Leave-Silo-Out (LSO), which aligns with the established terminology of cross-silo FL [8]. LSO serves as an estimator of the expected performance for free-riders. Simply put, LSO represents the expected prediction performance of an existing model collaboratively trained by other participants. Again, this is particularly relevant for small or newly established hospitals as a means to access high-quality predictive models. The training options are summarized in Fig. 1.

We evaluate the effectiveness of these approaches in the context of reducing unnecessary prostate cancer biopsies, a critical issue in modern urology. Our study highlights how different FL participation strategies impact prediction accuracy, investigating the potential practical benefits of federated learning for localized healthcare decision-making.

In prostate cancer (PCa) research, a key focus has been to incorporate a risk estimation model (or risk calculator) into the early detection pathway to reduce unnecessary biopsies, thereby preventing overdiagnosis and overtreatment. Many models have been proposed [27,25]; however, training and validating models that generalize well and accurately differentiate between benign/low-risk and high-risk PCa remains challenging. In this context, FL offers a solution by utilizing data from multiple institutions to create more representative and accurate models while preserving patient data privacy.

2. Related work

Realistic FL in Healthcare. Several previous studies have implemented realistic FL by distributing the training process across hospitals, using both images [31,33,26] and tabular data [2,39]. These studies have explored the potential of FL to address various medical challenges, such as breast density classification [31], COVID-19 detection [2,39], and prostate segmentation [33]. Notably, Pati et al. [26] conducted one of the largest real-world federated experiments, involving 71 medical institutions, thereby highlighting the challenges of evaluating FL model performance in practical settings [22].

Another approach to realistic FL involves partitioning data by origin, as seen in studies using hospital data from different regions [37,7,36,41,10,38]. Most FL studies evaluate global models using standard metrics like AUC and accuracy, while some focus on FL-specific metrics. For instance, an ensemble of local models from the four largest hospitals was used in [26].

Leave-Silo-Out. The term Leave-*Something*-Out has been used in medical research, primarily to describe non-federated cross-validation methods across hospitals. Examples include Leave-Site-Out [32], Leave-Hospital-Out [35], Leave-Institution-Out [6], and Leave-Source-Out [14]. These approaches focus on excluding one institution during model training and then using it as a test set to evaluate performance. In the context of FL, variations like Leave-One-(Institution)-Out [36] and Leave-Center-Out [15] have also been explored.

PCa Risk Calculator. The diagnosis of prostate cancer relies on prostate biopsies, which can expose men to significant complications, such as bleeding and severe infection [1]. Moreover, it is not uncommon for biopsies to yield benign results or reveal low-risk cancer [9]. Therefore, it is crucial to assess the individual's risk of significant prostate cancer before performing a biopsy to weigh the potential risks. In recent decades, several risk calculators have been developed for this purpose [21,28,5].

3. Methods

To reflect real-world data heterogeneity for cross-silo FL, we compiled 14 publicly available datasets from 10 countries. Each dataset is accompanied by a peer-reviewed article describing its data collection methods, clinical parameters, and procedures. This documentation allows for alignment and harmonization between institutions, ensuring consistency in the analysis. Fig. 2 illustrates the countries and cities where the participating hospitals are located.

3.1. Data

Each dataset includes key parameters relevant to PCa research: age in years, PSA in ng/ml, Prostate Volume (PV) in cm^3, Prostate Imaging–Reporting and Data System (PI-RADS) Score, and the presence of clinically significant PCa (csPCa) based on the Gleason Grade Group (GGG).

Although all datasets share the same features and variables, making their participation in FL mathematically feasible, we excluded the datasets from Türkiye and Finland from our study. The Turkish dataset has low feature diversity, as it includes only patients with PI-RADS 3. The Finnish dataset differs from other cohorts by including only men diagnosed with low-risk prostate cancer who are undergoing active surveillance. Ultimately, incorporating these datasets would be clinically unsound. Table 1 briefly summarizes the datasets used in this work. Since both Korea1 and Korea2 were collected at the same hospital, we combined them as *Korea (kr)*. A more detailed description of the datasets, their corresponding clinical tasks, the methods used to handle missing values, and the data preprocessing steps is provided in the supplementary materials.

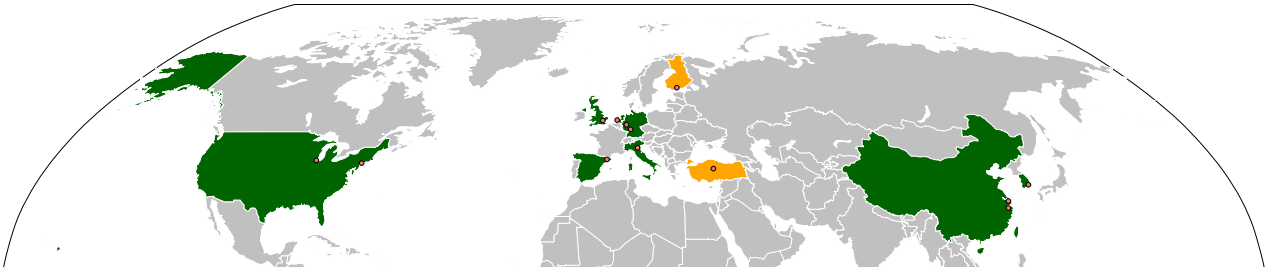


Fig. 2. World map showing the countries and cities where the hospitals and corresponding datasets are located. Datasets highlighted in orange are excluded from the experimental analysis due to clinical incompatibility.

Table 1
Clinical summary of the studied datasets.

Dataset	Incl	Study	Year	Patient #	Systematic Biopsy	Targeted Biopsy	PI-RADS	GGG ≥ 2
Spain (es)	✓	[19]	2021–2022	1640	✓	✓	2.0	✓
Korea1 (kr1)	✓	[13]	2014–2018	406	✓	✓	2.0	✓
Korea2 (kr2)	✓	[20]	2019–2021	593	✓	✓	2.1	✓
Germany1 (de1)	✓	[11]	2014–2019	774	✓	✓	2.1	✓
China2 (ch2)	✓	[23]	2018–2020	530	✓	✓	2.0	✓
China1(ch1)	✓	[42]	2013–2015	312	✓	✓	1.0	✓
USA2 (us2)	✓	[30]	2012–2014	310	✓	✓	2.0	✓
USA1 (us1)	✓	[18]	2014–2021	280	✓	✓	1.0/2.0/2.1	✓
Netherlands (nl)	✓	[42]	2013–2015	266	✓	✓	1.0	✓
Italy (it)	✓	[16]	2014–2016	218	✓	✓	2.0	✓
Germany2 (de2)	✓	[29]	2010–2012	162	✓	✓	2.0	✓
UK (gb)	✓	[29]	2010–2012	133	✓	✓	2.0	✓
Finland (fi)	✗	[4]	2006–2015	75	✓	✓	1.0	✓
Türkiye (tk)	✗	[34]	2019–2022	55	✓	✓	2.0	✓

3.1.1. Statistics of the datasets

This section provides an overview of the dataset statistics, highlighting heterogeneity across participating hospitals, and describes the steps taken to harmonize data across sites.

Features. To support cross-silo FL, we ensured that each dataset included a consistent set of features. Fig. 3a displays the distribution of the most influential features for prediction outcomes. Although the age distribution roughly follows a normal curve, PSA levels display extreme outliers, ranging from 0.7% to 13,300% of the mean. Additionally, PI-RADS scores vary significantly between hospitals.

To further explore differences in feature distributions, we applied t-Distributed Stochastic Neighbor Embedding (tSNE) to the datasets. However, since tSNE did not identify notable outliers within the studied datasets, the results are included only in the supplementary materials.

Labels. The primary label, csPCa, is defined as histopathological GGG ≥ 2 observed in prostate biopsies. The proportion of significant cancer cases varies between hospitals, as shown in Fig. 3b.

3.2. Model

We build on the work of Ettala et al. [5], which introduced a logistic regression (LR) model designed to prioritize the classification of patients with PI-RADS 3. Following their methodology, we adopt their recommended model architecture and coefficients as a *baseline* for our FL experiments, as shown in Fig. 5. This framework includes a non-linear spline transformation applied to PSA values to address non-linearity (see Fig. 3a). Additionally, a PI-RADS transformation assigns LR coefficients to scores of 3, 4, and 5, while scores below 3 are set to zero. This results in a model with a total of eight coefficients and an intercept. Although the original study adjusted PSA and PV values for patients based on 5-ARI medication, we exclude this factor from our analysis. This decision is based on its absence in the studied datasets and its minimal impact on overall model performance [5]. For model federation, we implement the Federated Averaging (FedAvg) algorithm [17]. A detailed description of the hyperparameters used can be found in the supplementary materials.

3.3. AUC estimators for training strategies

Let Z denote a $\mathcal{Z}$-valued random variable corresponding to a single datum, endowed with some unknown probability distribution. We decompose the variable and its outcome space as $Z = (X, Y, S)$ and $\mathcal{Z} = \mathcal{X} \times \{0, 1\} \times \mathcal{S}$, respectively, in which X takes its values from the set of possible feature vectors $\mathcal{X}$, Y is binary valued variable corresponding to the datum's class label and S is a nominally valued silo identity with $\mathcal{S}$ being the set of all silo identities. With $z = (x, y, z) \in \mathcal{Z}$, we denote a realization of Z. Finally, we use bold notation for denoting the d-dimensional $\mathcal{Z}^d$-valued random vectors variables $\boldsymbol{Z}$ endowed with the dth order product probability distribution of Z and analogously $\boldsymbol{z} \in \mathcal{Z}^d$ for its realizations.

Let $C_{n,s} \subset \mathcal{Z}^d$ be the subset of data sequences of length d that consists of data from $K + 1$ distinct silos and at least $5/4n$ data originating from silo s in particular. Moreover, let t be a function, whose value $t(\boldsymbol{z}, n, s)$ is the AUC of a federated learning model trained on all data from the all silos present in $\boldsymbol{z}$ except only the first n data from silo s in $\boldsymbol{z}$, tested on the remaining $1/4n$ data from the silo s not used for training. Then,

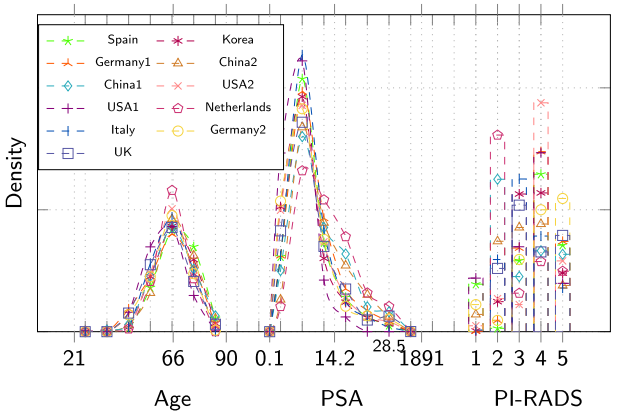
$$\theta^{\mathrm{FL}}_{n,s} = \mathrm{E}[t(\boldsymbol{Z}, n, s) \mid \boldsymbol{Z} \in C_{n,s}]$$

is the FL approach's expected AUC on new data from silo s, given that the FL model is trained with exactly n other data from the same silo and $d - 5/4n$ data from any other K distinct silos. Analogously,

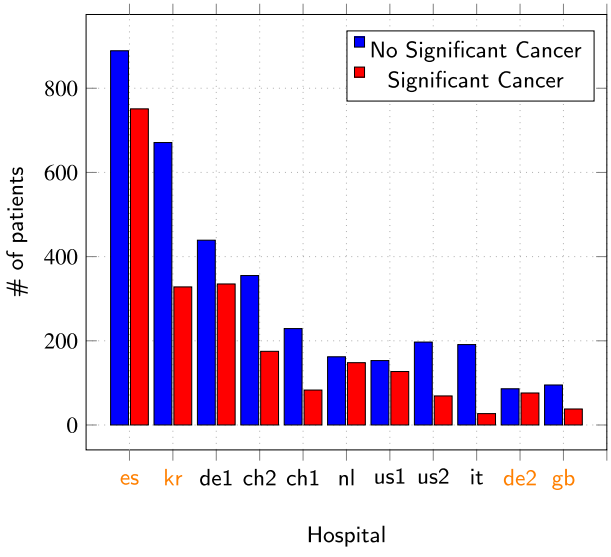
$$\theta^{\mathrm{FL}}_{n} = \mathrm{E}[t(\boldsymbol{Z}, n, S) \mid \boldsymbol{Z} \in C_{n,S}]\,,$$

is the same quantity without fixing the test silo. As special cases with $n = 0$, we obtain the analogous estimands for the FR approach, namely $\theta^{\mathrm{FR}} = \theta^{\mathrm{FL}}_{0}$ and $\theta^{\mathrm{FR}}_{s} = \theta^{\mathrm{FL}}_{0,s}$.

The statistic t itself can be considered as an estimator of $\theta^{\mathrm{FL}}_{n,s}$, hence its value on a given $\boldsymbol{z} \in C_{n,s}$ provides an estimate for it. A standard approach for reducing the variance of the estimator is to average it over all reorderings of the data sequence that are conformable with its domain. That is,



(a) Features Distribution



(b) Label Distribution

Fig. 3. Datasets statistics. For (a), the minimum, maximum, and mean values are illustrated. While all features are linearly spaced, the final bin for PSA ranges from 28.5 to 1891. For (b), the datasets highlighted in orange lack raw GGG scores and only include binary 0 (blue) or 1 (red) labels for csPCa, preventing label recalibration.

$$\widehat{\theta}^{\mathrm{FL}}_{n,s}(\boldsymbol{z}) = \sum_{\pi \in \Pi_{n,s}} t(\pi \cdot \boldsymbol{z}, n, s)$$

where $\Pi_{n,s}$ is the subset of all bijections $\pi : \{1, \ldots, d\} \to \{1, \ldots, d\}$ such that $\pi \cdot \boldsymbol{z} \in C_{n,s}$, and $\pi \cdot \boldsymbol{z}$ denotes the data sequence, whose entries are reordered from $\boldsymbol{z}$ according to π. Similarly, a reduced variance estimator for θ^{FL}_n can be obtained as

$$\widehat{\theta}^{\mathrm{FL}}_{n}(\boldsymbol{z}) = \sum_{s} \widehat{\theta}^{\mathrm{FL}}_{n,s}(\boldsymbol{z})$$

where the sum is taken over all silo identities that have at least $5/4n$ data in $\boldsymbol{z}$. These are what we refer to as the leave-silo-partly-out (LSPO) estimators and for the FR special cases the leave-silo-out (LSO) estimators.

In practice, averaging over all conformable reorderings is infeasible, and hence one usually resorts to using only a representative subset of them. Here, for estimators of $\theta^{\mathrm{FL}}_{n,s}$, we average over 10000 reorderings. The complete pseudocode of the LSPO estimator can be seen in Algorithm 1, where $\{\boldsymbol{z}_k\}_{k \in S \setminus \{s\}}$ represents data from K distinct silos. The LSO estimator can also be derived from it by using all data from $\boldsymbol{z}_s$ for testing and none for training.

Algorithm 1 LSPO Estimation in Mathematical Notation.

Input: $\{\boldsymbol{z}_s\}_{s \in S}$, number of repetitions R, number of federated rounds T
for each $s \in S$ **do**
  for $r = 1$ to R **do**
    $\pi_r \leftarrow$ randomly from $\Pi_{n,s}$
    $\boldsymbol{z}_s^{\mathrm{train}} \leftarrow n$ first data from $(\pi_r \cdot \boldsymbol{z})_s$
    $\boldsymbol{z}_s^{\mathrm{test}} \leftarrow 20\%$ last data from $(\pi_r \cdot \boldsymbol{z})_s$
    for each $k \in S \setminus \{s\}$ **do**
      $\boldsymbol{z}_k^{\mathrm{train}} \leftarrow 80\%$ first data from $(\pi_r \cdot \boldsymbol{z})_k$
    Initialize global model parameters $\boldsymbol{w}_0$
    for $t = 1$ to T **do**
      for each $k \in S \setminus \{s\}$ **do**
        Train local model on $\boldsymbol{z}_k^{\mathrm{train}}$
        Compute local update $\Delta \boldsymbol{w}_k$
      Train local model on $\boldsymbol{z}_s^{\mathrm{train}}$
      Compute local update $\Delta \boldsymbol{w}_s$
      Aggregate: $\boldsymbol{w}_t \leftarrow \mathcal{A}\left(\{\Delta \boldsymbol{w}_k\}, \Delta \boldsymbol{w}_s\right)$
    Compute $t(\pi_r \cdot \boldsymbol{z}, n, s)$ as AUC evaluated on $\boldsymbol{z}_s^{\mathrm{test}}$
  Compute $\widehat{\theta}^{\mathrm{FL}}_{n,s}(\boldsymbol{z}) = \frac{1}{R} \sum_{r=1}^{R} t(\pi_r \cdot \boldsymbol{z}, n, s)$
Output: $\{\widehat{\theta}^{\mathrm{FL}}_{n,s}(\boldsymbol{z})\}_{s \in S}$,

4. Results

In this study, we explore three real-world strategies for hospitals seeking to obtain a diagnostic model: Local Training, FL Training, or Free-riding (FR). All modeling decisions were made on the *training* sets for each repetition of all the experiments done in this work. Our approach begins by training models on local data using an 80/20 train-test split. These locally trained models are then evaluated on all other test datasets, creating a federated evaluation matrix shown in the upper part of Fig. 4. It displays the results averaged over 10,000 Monte Carlo simulations, each with random train/test splits. The rows correspond to the hospitals that trained the models, while the columns represent the datasets used for testing. The datasets are sorted in descending order by patient count (see Fig. 3b). We report two aggregation metrics, AV-H and AV-W. AV-H represents the average test performance for each hospital (row). AV-W calculates AV-H weighted by the number of patients in each dataset. The highlighted diagonal shows the local accuracies achieved by each hospital, with AV-DIAG representing the average performance of locally trained models. This serves as a baseline metric, with an AUC of 84.0, indicating the expected accuracy of a locally trained model.

To compare with Local Training, we evaluated models trained using multi-hospital learning. The results, averaged over 10,000 Monte Carlo simulations with random train/test splits, are presented in the lower matrix of Fig. 4. As an optimal baseline, a centralized model (row 1) is trained on the combined training data from all sites and tested on the respective 20% test parts. FL results using the FedAvg algorithm (row 2) are similarly derived, where models are trained on hospital training parts and evaluated on their corresponding test data. FR performance is estimated with LSO by sequentially excluding one institution during the training phase and testing the resulting model on the entire dataset of the excluded hospital (rows 3–13). AV-H, AV-W, and AV-DIAG are defined as before, with AV-DIAG in this context reflecting the *average* performance of FR (LSO) on local data. We focus on comparing the two diagonals, ranked from highest to lowest based on the number of local patients.

As expected, the centralized model achieves the highest overall performance with an average AUC of 84.7. When comparing the three practical training strategies (Local, FL, and FR), FL achieves the highest performance with an average AUC of 84.2 (highlighted in blue). However, both Local Training (AV-DIAG = 84.0) and FR (AV-DIAG = 83.9) demonstrate comparable results. Notably, the advantages of FL over FR diminish for hospitals with smaller datasets (fewer than 200 patients). For institutions like Italy, Germany2, and UK, FL provides only

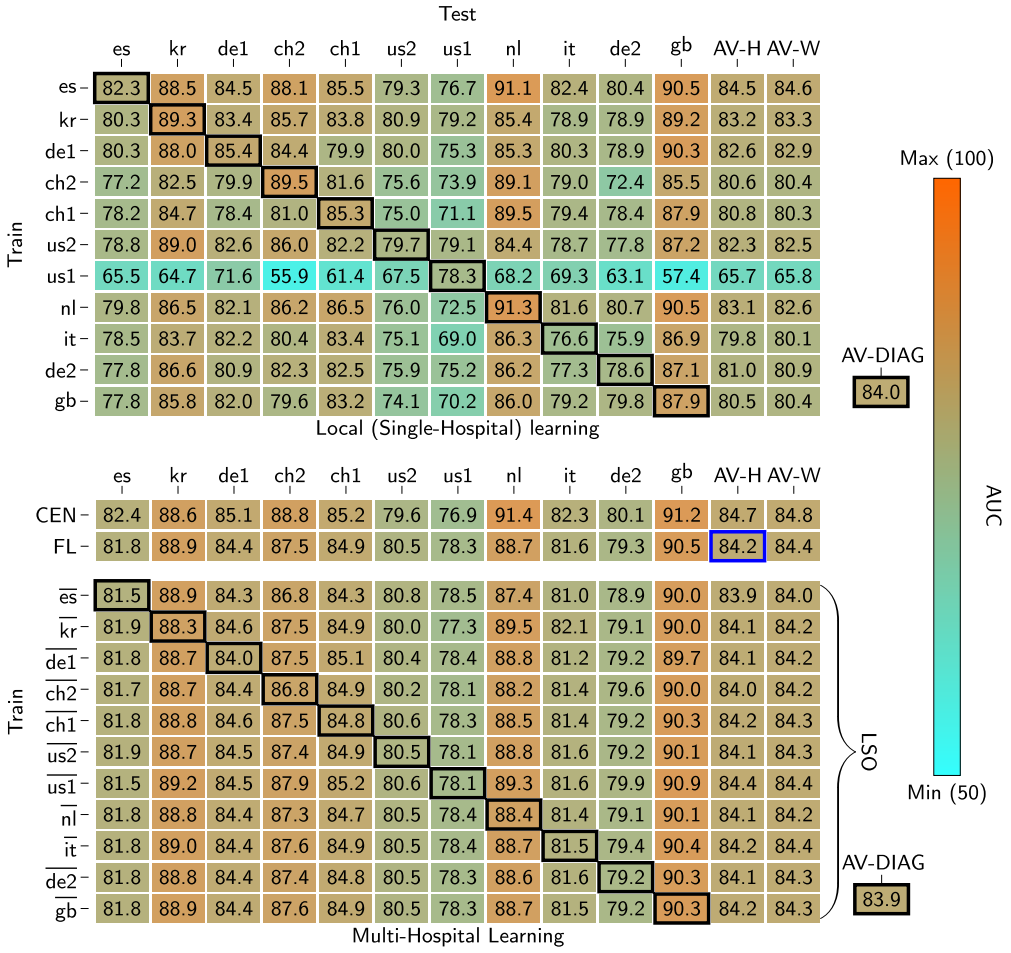


Fig. 4. Heatmap Matrices of Single-Hospital Learning (top) and Multi-Hospital Learning (bottom). Rows represent the training hospitals, and columns indicate those used for testing. Datasets are labeled by their country alpha-2 codes, with code marking silos excluded from federated training. AV-H is the average test performance across testing hospitals (row), while AV-W provides the patient-weighted average. AV-DIAGs show the average result along the highlighted diagonals. CEN denotes results from centralized training, and FL indicates the federated model trained across all participating hospitals. Datasets are arranged in descending order by sample size.

a marginal improvement over FR. Conversely, for larger silos such as Spain, Korea, and Germany1, FL significantly underperforms compared to Local Training. These observations suggest that FL may not always be the most effective strategy for all hospitals. USA1 presents a particularly interesting case. Its locally trained model performs poorly, especially when cross-evaluated with data from other sites. However, under the FR setup, where USA1 is excluded during training, the resulting global model achieves an AUC of 84.4, outperforming FL. This outcome points to specific challenges in the USA1 local dataset, such as variations in data quality or other unmeasured factors.

Although Local Training and FR produce similar overall results, their effectiveness in local data varies significantly depending on hospital dataset size. A comparison of the two matrices reveals that Local Training excels for larger datasets (top-left of the highlighted diagonal), whereas FRis more effective for smaller datasets (bottom-right of the highlighted diagonal). It appears that hospitals with larger population are likely to achieve the best diagnostic results using Local Training, while those with limited data gain more from being a free-rider. Although Local Training performs well for local patient outcomes, its cross-evaluation results on data from other sites are understandably inferior. Analyzing the *off-diagonal* values in the matrices shows that, on average, FR outperforms Local Training by 4.1 AUC points in cross-evaluation, underscoring its advantage in multi-hospital scenarios.

To explore these findings further, we analyze how dataset size impacts the effectiveness of each training strategy. For each dataset, we randomly fixed the test set size at 20% of the dataset, varied the training size, and compared three possible training methods alongside the baseline model. Since the baseline and FR models do not involve the hospital in question in training, their performance remains constant regardless of training data. To ensure robust results, we performed 1,000 simulations. The complete pseudocode of the pipeline can be found in Appendix. The first 13 plots of Fig. 5 (Spain through UK) display how dataset size affects AUC performance for each hospital. The final plot (Average Difference) summarizes the average AUC differences between Local Training and FR Local and FL, as well as between FL and FR, across all sites. For this aggregated graph, only hospitals with a sufficient number of patients were included for each sample size. The analysis is limited to 400 samples, as the number of contributing hospitals drops to three beyond this threshold, reducing the representativeness of the results.

The results indicate that AUC for Local Training improves steadily as the number of samples increases, often exceeding FL performance once a hospital's dataset is sufficiently large. Although FL performance also improves with larger sample sizes, its rate of improvement is slower. The final graph demonstrates that when sample sizes reach approximately 250, Local Training consistently outperforms Multi-hospital Training, suggesting it may be the preferred strategy for institutions with enough data to effectively diagnose local patients. Similarly, FL begins to noticeably outperform FR only when hospitals have a sufficient number of samples, approximately 250. However, with too few samples, a hospital's contribution does not significantly enhance the global model, and by the time FL outperforms FR, the local model has already become the most effective option for local patients. Considering the trade-off between setup complexity, privacy, and benefit, the "optimal" solution for a hospital appears to be a free-rider until a sufficient number of local

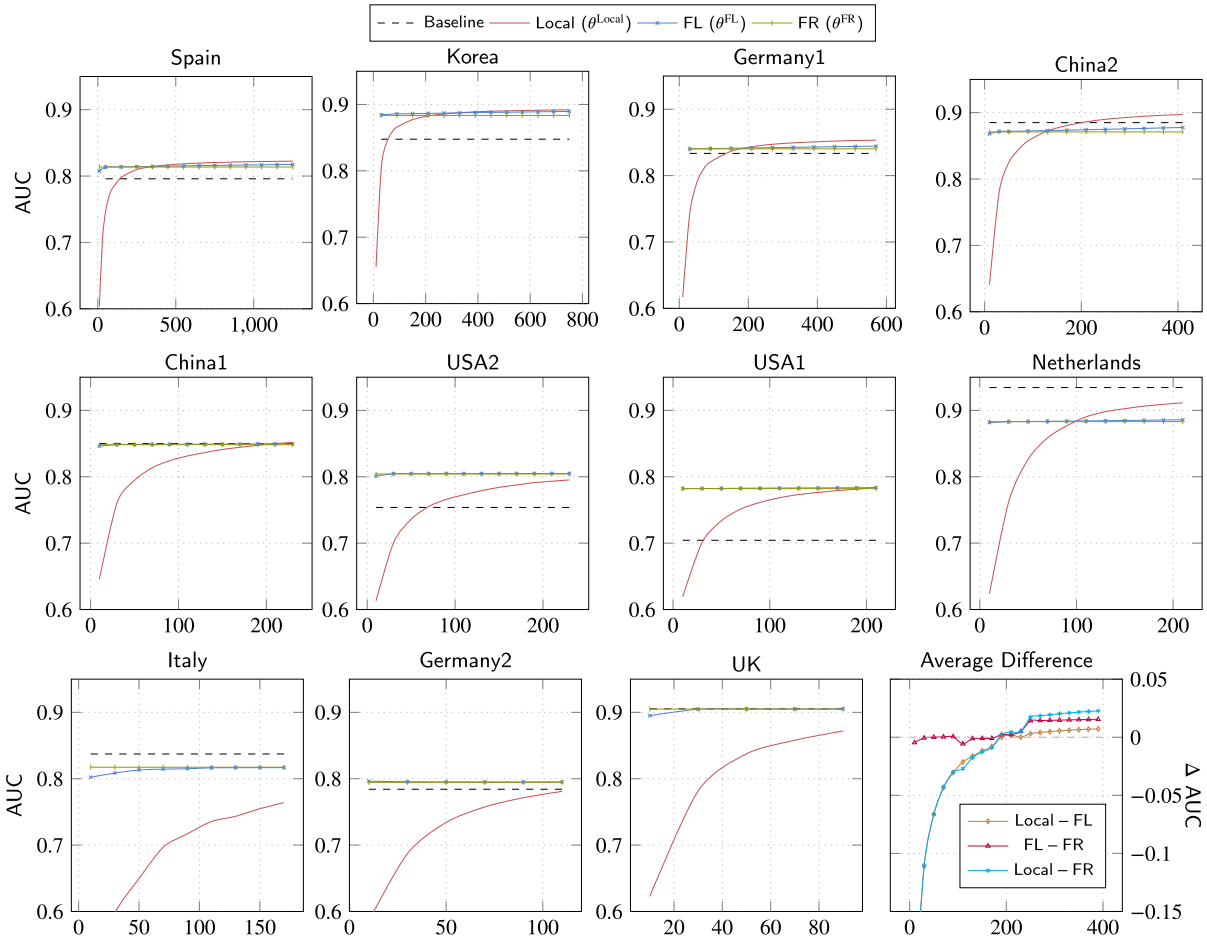


Fig. 5. Effect of Varying Training Dataset Size on AUC for Local Test Data. Baseline and Free-riding (FR) curves are horizontal as these hospitals do not participate in training. 'Average Difference' shows the difference between training methods. Only hospitals with a sufficient number of patients contribute to the average.

samples is available, which in our use case is approximately 250. Once this threshold is reached, the hospital can transition to training its own local model without ever contributing to the federated model.

5. Discussion

Our findings suggest that no single federated strategy is ideal for all hospitals or scenarios. Institutions with substantial patient data may find training a local model more viable, avoiding the complexity of federated infrastructure. Conversely, smaller hospitals might prefer the free-rider approach due to minimal AUC loss compared to federated training and privacy concerns associated with sharing local model updates. As a result, participation in federated training may never be the optimal choice for these institutions.

However, the availability of high-quality off-the-shelf models depends on large, self-sufficient institutions contributing to the federated model's development. If too many contributors opt out, the global model could degrade and become unusable. A potentially viable setup might involve a fixed number of large hospitals maintaining a federated model while simultaneously managing their own internal models. This federated model could then be shared – either freely or with financial incentives – with smaller hospitals lacking the data to train their own models. Importantly, when local model performance does not significantly surpass that of federated model, the latter may be preferred due to its superior generalizability across hospitals.

Questions about learning curves relative to the number of participating hospitals and the practical implementation of such setups remain open. We leave these considerations for future research.

Real-world Heterogeneity. The methodological differences between the studied datasets, such as variations in biopsy procedures, PI-RADS versions, and MRI equipment, could be seen as a limitation due to the resulting heterogeneity. However, these differences are a strength of this study, as they reflect the reality of a new hospital joining a federated consortium with its unique protocols. This heterogeneity enables us to explore the most effective training strategies under real-world conditions.

Clinical considerations. Although the LR models used in this study include only eight features, clinical practice demonstrates that even relying solely on the PI-RADS score can yield reasonable AUC. In practice, patients with PI-RADS scores of 4 and 5 are typically referred for biopsy, while those with scores of 1 and 2 are not. The greatest challenge lies in accurately classifying men with PI-RADS 3, where all eight features contribute significantly to improving outcomes for this group.

The original risk calculator paper [5] used GGG ≥ 3 as the threshold for defining csPCa. However, all datasets in this study adopt a different cutoff, GGG ≥ 2, which slightly alters the clinical question addressed by the trained models. Since raw GGG scores were unavailable for some hospitals (see Fig. 3b), recalibrating the csPCa definition was not feasible. To ensure comparability across all datasets, we adhered to this common definition. Despite the baseline model being tuned for a slightly different task, its performance was only marginally worse compared to other methods.

Limitations. The main limitation of this study is that it was conducted in a simulated setting without actual data distribution across distinct hospitals. While such physical separation is a critical aspect of real-world cross-hospital deployments, we believe the findings of this paper are still

applicable and translatable to practical federated learning scenarios. Another limitation is that we tested our approach on only the problem of PCa. Naturally, it would be valuable to evaluate it on other diseases and datasets as well. However, it is rare to find large publicly available collections of medical datasets that genuinely originate from distinct hospitals. This scarcity makes our contribution of sourcing open-access PCa datasets and publishing them even more valuable. A further limitation of our work is the lack of discussion around the privacy concerns inherent to federated learning. Although FL avoids direct data sharing, it still involves the exchange of model updates, which alone may not be sufficient to guarantee data privacy. Attacks such as model inversion attacks [40] can potentially lead to the inference of sensitive patient information. Therefore, additional techniques, such as differential privacy [3], may be necessary to enhance protection. We leave exploring privacy-preserving strategies in federated learning to future work.

6. Conclusion

In this work, we gathered geographically diverse datasets on prostate cancer and evaluated practical FL training strategies. We show that, despite significant improvements in cross-hospital evaluation, federated learning does not always outperform alternative training strategies for local disease predictions. This is especially true given the practical complexities, privacy concerns, and regulatory challenges involved in establishing federated learning connections. The optimal strategy for each hospital is likely to be situation-specific.

Summary table

What was already known?

- FL medical prediction models can achieve results comparable to centralized learning.
- It is challenging to evaluate FL medical prediction models in *real-world* settings.
- The “free-rider” mode allows hospitals to obtain a predictive model in federated settings.
- Prostate cancer risk calculators can be utilized to reduce unnecessary biopsies.

What does this study add to our knowledge?

- We propose a realistic framework for evaluating FL prediction models designed for medical tasks in cross-hospital settings, outlining practical strategies for institutions to develop high-performing diagnostic models for their local patients.
- We introduce and systematize a *Leave-Silo-Out* evaluation strategy for FL, which allows estimating how effectively hospitals can leverage federated models when not actively participating in training.
- We compile and publish a list of openly available prostate cancer datasets sourced from peer-reviewed journals. Additionally, we provide the code to process these datasets and implement our benchmarks.
- To the best of our knowledge, this is the first study to implement and analyze cross-silo FL for prostate cancer, with a focus on reducing unnecessary biopsies.

CRediT authorship contribution statement

Andrei Kazlouski: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Ileana Montoya Perez:** Writing – review & editing, Methodology. **Faiza Noor:** Writing – review & editing, Investigation. **Mikael Högerman:** Writing – review & editing, Formal analysis. **Otto Ettala:** Writing – review & editing, Formal analysis, Data curation. **Tapio Pahikkala:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Antti Airola:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of competing 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.

Acknowledgement

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).

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2025.106046.

Data availability

The code, along with all data used in this paper to ensure full reproducibility of the results, is openly accessible at https://github.com/phaseivai/Towards-Practical-FL/.

References

[1] M. Borghesi, H. Ahmed, R. Nam, E. Schaeffer, R. Schiavina, S. Taneja, W. Weidner, S. Loeb, Complications after systematic, random, and image-guided prostate biopsy, Eur. Urol. 71 (2017) 353–365.
[2] I. Dayan, H.R. Roth, A. Zhong, A. Harouni, A. Gentili, A.Z. Abidin, A. Liu, A.B. Costa, B.J. Wood, C.S. Tsai, et al., Federated learning for predicting clinical outcomes in patients with COVID-19, Nat. Med. 27 (2021) 1735–1743.
[3] C. Dwork, Differential privacy, in: International Colloquium on Automata, Languages, and Programming, Springer, 2006, pp. 1–12.
[4] J.T. Eineluoto, P. Järvinen, A. Kenttämies, T.P. Kilpeläinen, H. Vasarainen, K. Sandeman, A. Erickson, T. Mirtti, A. Rannikko, Repeat multiparametric MRI in prostate cancer patients on active surveillance, PLoS ONE 12 (2017) e0189272.
[5] O. Ettala, I. Jambor, I.M. Perez, M. Seppänen, A. Kaipia, H. Seikkula, K.T. Syvänen, P. Taimen, J. Verho, A. Steiner, 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), BMJ Open 12 (2022) e053118.
[6] T. Hollon, C. Jiang, A. Chowdury, M. Nasir-Moin, A. Kondepudi, A. Aabedi, A. Adapa, W. Al-Holou, J. Heth, O. Sagher, et al., Artificial-intelligence-based molecular classification of diffuse gliomas using rapid, label-free optical imaging, Nat. Med. 29 (2023) 828–832.
[7] M. Jiang, H.R. Roth, W. Li, D. Yang, C. Zhao, V. Nath, D. Xu, Q. Dou, Z. Xu, Fair federated medical image segmentation via client contribution estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16302–16311.
[8] P. Kairouz, H.B. McMahan, B. Avent, A. Bellet, M. Bennis, A.N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, et al., Advances and open problems in federated learning, Found. Trends Mach. Learn. 14 (2021) 1–210.
[9] V. Kasivisvanathan, A.S. Rannikko, M. Borghi, V. Panebianco, L.A. Mynderse, M.H. Vaarala, A. Briganti, L. Budäus, G. Hellawell, R.G. Hindley, et al., MRI-targeted or standard biopsy for prostate-cancer diagnosis, N. Engl. J. Med. 378 (2018) 1767–1777.
[10] R. Kerkouche, G. Acs, C. Castelluccia, P. Genevès, 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.
[11] M. Klingebiel, C. Arsov, T. Ullrich, M. Quentin, R. Al-Monajjed, D. Mally, L. Sawicki, A. Hiester, I. Esposito, P. Albers, et al., Data on the detection of clinically significant prostate cancer by magnetic resonance imaging (MRI)-guided targeted and systematic biopsy, Data Brief 45 (2022) 108683.
[12] T.T. Kuo, A. Pham, Detecting model misconducts in decentralized healthcare federated learning, Int. J. Med. Inform. 158 (2022) 104658.
[13] S.S. Lee, D.H. Lee, W.H. Song, J.K. Nam, J.Y. Han, H.J. Lee, T.U. Kim, S.W. Park, Usefulness of bi-parametric magnetic resonance imaging with b = 1,800 s/mm^2 diffusion-weighted imaging for diagnosing clinically significant prostate cancer, World J. Men’s Health 38 (2020) 370.

[14] T. Leinonen, D. Wong, A. Vasankari, A. Wahab, R. Nadarajah, M. Kaisti, A. Airola, Empirical investigation of multi-source cross-validation in clinical ECG classification, Comput. Biol. Med. 183 (2024) 109271.
[15] A. Linardos, K. Kushibar, S. Walsh, P. Gkontra, K. Lekadir, Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease, Sci. Rep. 12 (2022) 3551.
[16] E. Martorana, G.M. Pirola, M. Scialpi, S. Micali, A. Iseppi, L.R. Bonetti, S. Kaleci, P. Torricelli, G. Bianchi, Lesion volume predicts prostate cancer risk and aggressiveness: validation of its value alone and matched with prostate imaging reporting and data system score, BJU Int. 120 (2017) 92–103.
[17] B. McMahan, E. Moore, D. Ramage, S. Hampson, B.A. y Arcas, Communication-efficient learning of deep networks from decentralized data, in: Artificial Intelligence and Statistics, in: PMLR, 2017, pp. 1273–1282.
[18] J. Meza, R. Babajide, R. Saoud, J. Sweis, J. Abelleira, I. Helenowski, B. Jovanovic, S. Eggener, F.H. Miller, J.M. Horowitz, et al., Assessing the accuracy of multiparametric MRI to predict clinically significant prostate cancer in biopsy naïve men across racial/ethnic groups, BMC Urol. 22 (2022) 107.
[19] J. Morote, N. Picola, J. Muñoz-Rodriguez, N. Paesano, X. Ruiz-Plazas, M.V. Muñoz-Rivero, A. Celma, G. García-de Manuel, B. Miró, J.M. Abascal, 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, World J. Men's Health 42 (2024) 441.
[20] J.K. Nam, W.H. Song, S.S. Lee, H.J. Lee, T.U. Kim, S.W. Park, Impact of ultrasonographic findings on cancer detection rate during magnetic resonance image/ultrasonography fusion-targeted prostate biopsy, World J. Men's Health 41 (2023) 743.
[21] R.K. Nam, M.W. Kattan, J.L. Chin, J. Trachtenberg, R. Singal, R. Rendon, L.H. Klotz, L. Sugar, C. Sherman, J. Izawa, et al., Prospective multi-institutional study evaluating the performance of prostate cancer risk calculators, J. Clin. Oncol. 29 (2011) 2959–2964.
[22] T.M. Nguyen, K.L. Poh, S.L. Chong, J.H. Lee, FedDSS: a data-similarity approach for client selection in horizontal federated learning, Int. J. Med. Inform. 192 (2024) 105650.
[23] J.F. Pan, R. Su, J.Z. Cao, Z.Y. Zhao, D.W. Ren, S.Z. Ye, R.D. Huang, Z.L. Tao, C.L. Yu, J.H. Jiang, et al., Modified predictive model and nomogram by incorporating pre-biopsy biparametric magnetic resonance imaging with clinical indicators for prostate biopsy decision making, Front. Oncol. 11 (2021) 740868.
[24] E. Parliament, General data protection regulation GDPR, Online, https://gdpr-info.eu/, 2016. (Accessed 25 November 2024).
[25] H.D. Patel, S. Remmers, J.L. Ellis, E.V. Li, M.J. Roobol, A.M. Fang, P. Davik, S. Rais-Bahrami, A.B. Murphy, A.E. Ross, et al., Comparison of magnetic resonance imaging–based risk calculators to predict prostate cancer risk, JAMA Netw. Open 7 (2024) e241516.
[26] S. Pati, U. Baid, B. Edwards, M. Sheller, S.H. Wang, G.A. Reina, P. Foley, A. Gruzdev, D. Karkada, C. Davatzikos, et al., Federated learning enables big data for rare cancer boundary detection, Nat. Commun. 13 (2022) 7346.
[27] N. Pereira-Azevedo, J.F. Verbeek, D. Nieboer, C.H. Bangma, M.J. Roobol, Head-to-head comparison of prostate cancer risk calculators predicting biopsy outcome, Transl. Androl. Urol. 7 (2018) 18.
[28] M. Peters, D. Eldred-Evans, P. Kurver, U.G. Falagario, M.J. Connor, T.T. Shah, J.J. Verhoeff, P. Taimen, H.J. Aronen, J. Knaapila, 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, Eur. Urol. 82 (2022) 559–568.
[29] J.P. Radtke, F. Giganti, M. Wiesenfarth, A. Stabile, J. Marenco, C. Orczyk, V. Kasivisvanathan, J.N. Nyarangi-Dix, V. Schütz, S. Dieffenbacher, 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, PLoS ONE 14 (2019) e0221350.
[30] A.R. Rastinehad, N. Waingankar, B. Turkbey, O. Yaskiv, A.M. Sonstegard, M. Fakhoury, C.A. Olsson, D.N. Siegel, P.L. Choyke, E. Ben-Levi, et al., Comparison of multiparametric MRI scoring systems and the impact on cancer detection in patients undergoing mr us fusion guided prostate biopsies, PLoS ONE 10 (2015) e0143404.
[31] H.R. Roth, K. Chang, P. Singh, N. Neumark, W. Li, V. Gupta, S. Gupta, L. Qu, A. Ihsani, B.C. Bizzo, et al., Federated learning for breast density classification: a real-world implementation, in: Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning: Second MICCAI Workshop, DART 2020, and First MICCAI Workshop, DCL 2020, Springer, 2020, pp. 181–191.
[32] M. Rozycki, T.D. Satterthwaite, N. Koutsouleris, G. Erus, J. Doshi, D.H. Wolf, Y. Fan, R.E. Gur, R.C. Gur, E.M. Meisenzahl, et al., Multisite machine learning analysis provides a robust structural imaging signature of schizophrenia detectable across diverse patient populations and within individuals, Schizophr. Bull. 44 (2018) 1035–1044.
[33] K.V. Sarma, S. Harmon, T. Sanford, H.R. Roth, Z. Xu, J. Tetreault, D. Xu, M.G. Flores, A.G. Raman, R. Kulkarni, et al., Federated learning improves site performance in multicenter deep learning without data sharing, J. Am. Med. Inform. Assoc. 28 (2021) 1259–1264.
[34] S. Senel, K. Ceviz, Y. Kasap, S. Tastemur, E. Olcucuoglu, E. Uzun, M.E. Polat, A. Koudonas, F. Sarialtin, Efficacy of plasma atherogenic index in predicting malignancy in the presence of prostate imaging–reporting and data system 3 (PI-RADS 3) prostate lesions, Int. Urol. Nephrol. 55 (2023) 255–261.
[35] J.K. Shade, A.N. Doshi, E. Sung, D.M. Popescu, A.S. Minhas, N.A. Gilotra, K.N. Aronis, A.G. Hays, N.A. Trayanova, Real-time prediction of mortality, cardiac arrest, and thromboembolic complications in hospitalized patients with COVID-19, JACC Adv. 1 (2022) 100043.
[36] M.J. Sheller, B. Edwards, G.A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R.R. Colen, et al., Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data, Sci. Rep. 10 (2020) 12598.
[37] M.J. Sheller, G.A. Reina, B. Edwards, J. Martin, S. Bakas, Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation, in: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Springer, 2019, pp. 92–104.
[38] J. Ogier du Terrail, S.S. Ayed, E. Cyffers, F. Grimberg, C. He, R. Loeb, P. Mangold, T. Marchand, O. Marfoq, E. Mushtaq, et al., Flamby: datasets and benchmarks for cross-silo federated learning in realistic healthcare settings, Adv. Neural Inf. Process. Syst. 35 (2022) 5315–5334.
[39] A. Vaid, S.K. Jaladanki, J. Xu, S. Teng, A. Kumar, S. Lee, S. Somani, I. Paranjpe, J.K. De Freitas, T. Wanyan, et al., Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach, JMIR Med. Inform. 9 (2021) e24207.
[40] X. Wu, M. Fredrikson, S. Jha, J.F. Naughton, A methodology for formalizing model-inversion attacks, in: 2016 IEEE 29th Computer Security Foundations Symposium (CSF), IEEE, 2016, pp. 355–370.
[41] R. Yan, L. Qu, Q. Wei, S.C. Huang, L. Shen, D.L. Rubin, L. Xing, Y. Zhou, Label-efficient self-supervised federated learning for tackling data heterogeneity in medical imaging, IEEE Trans. Med. Imaging 42 (2023) 1932–1943.
[42] K. Zhang, R. Chen, A.R. Alberts, G. Zhu, Y. Sun, M.J. Roobol, 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, Prostate Int. 7 (2019) 96–101.