# Feasibility of Fully Data-Driven Federated Learning on Large Observational Health Data

**Egill A. Fridgeirsson[1], Jenna M. Reps[1,2]**

**[1]Department of Medical Informatics Erasmus University Medical Center, Rotterdam, the Netherlands**

**[2]JnJ, Raritan, NJ, USA**

### Background

Large routinely-collected claims and EHR data enable the development of prognostic models for many clinical decisions. Because these databases are high-dimensional and extremely sparse, L1-penalised logistic regression is widely regarded as the best-performing approach [1,2]. Privacy regulations, however, often prevent pooling patient-level records. Federated learning (FL) overcomes this by transferring only aggregated model parameters. Most healthcare FL demonstrations use one-shot or few-shot averaging, require a small pre-selected feature set, and can keep a human “in the loop” due to few communication rounds. Google’s federated dual averaging (FDA) algorithm [3] generalizes L1 optimization to a multi-round FL setting, theoretically allowing thousands of candidate predictors to be considered. Yet its practical feasibility on real-world observational health data has never been quantified. Here we evaluate the predictive performance and computational burden of FDA-based FL compared with conventional, centrally trained (pooled) L1 logistic regression across multiple clinical prediction tasks and feature granularities.

### Methods

Three prediction tasks were extracted from the Integrated Primary Care Information (IPCI) to the OMOP common data model:

- Predicting dementia within five years in patients aged 55-84 after a healthcare visit.
- Predict lung cancer in next three years in adults after a healthcare visit.
- Predict readmission within 30 days after hospital discharge.

For each task we constructed two alternative predictor sets:

- Regular features – age, sex, and a binary indicator for conditions recorded in the last year
- Phenotype features – age, sex, plus 63 clinically curated phenotypes that use common condition and drug concepts from the last year [4].

We split the datasets into five equal parts, stratified by outcome. Then we set up a simulated federated learning environment where a server process creates five worker processes (see Figure 1). Only the worker processes have access to non-aggregated data and each one only sees one split of the data. The server process receives aggregated update information from the clients, updates its state and then sends the new server state down to the worker processes. The client and server update functions are both implemented in c++ using RcppEigen.

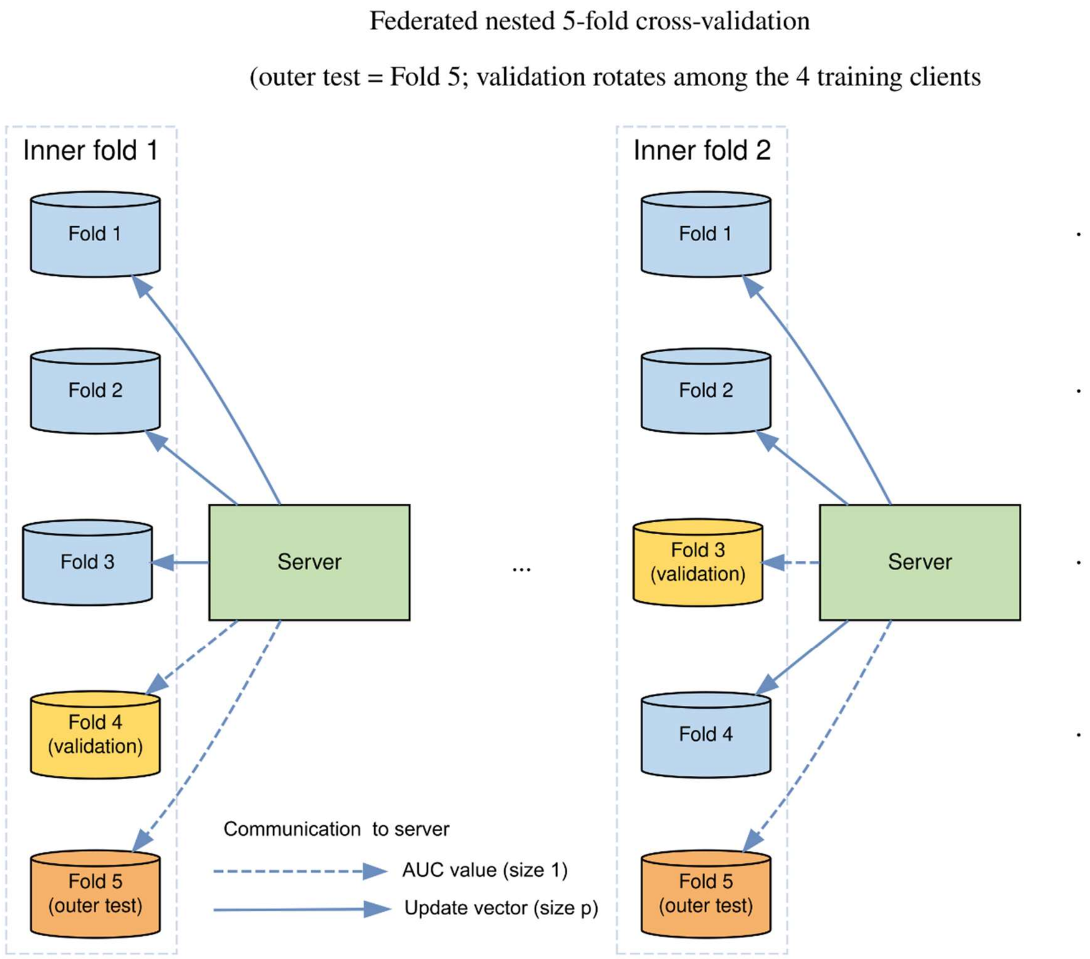


**Figure 1:** Single iteration of the outer loop of the nested cross validation.

We use five-fold nested cross-validation to evaluate the algorithms where each part is a fold. The model is developed on four parts. We use cross validation with four parts (develop on three and evaluate on the fourth) to select the best hyperparameter strength. We use Cyclops's auto search to select the best penalty strength to avoid many rounds of hyperparameter tuning [5]. Finally, once we identified the best hyperparmeter we evaluate the performance on the fifth part. Then we repeat this five times, with each part as the evaluation/test part. The centrally trained (pooled) follows the same approach but uses the PatientLevelPrediction framework [6].

We use the area under the receiver operating characteristics curve (AUROC) to evaluate performance as well as total runtime.

Analyses were executed on Ubuntu 24.04 LTS (Linux kernel 6.8.0) on a server with 2x Intel® Xeon® Platinum 8462Y+ (128 logical cores), 1.0TB RAM. Code was written in R v4.4.1 with Cyclops v3.4.1, PatientLevelPrediction v6.4.1, and RcppEigen v0.3.4; C++17 code compiled with g++ 13.3, linked against OpenBLAS v0.3.26. Federated learning was simulated on a single host with one coordinator and five worker processes; each worker used 1 thread. Wall-clock runtime was recorded as Sys.time() start/stop around training.

## Results

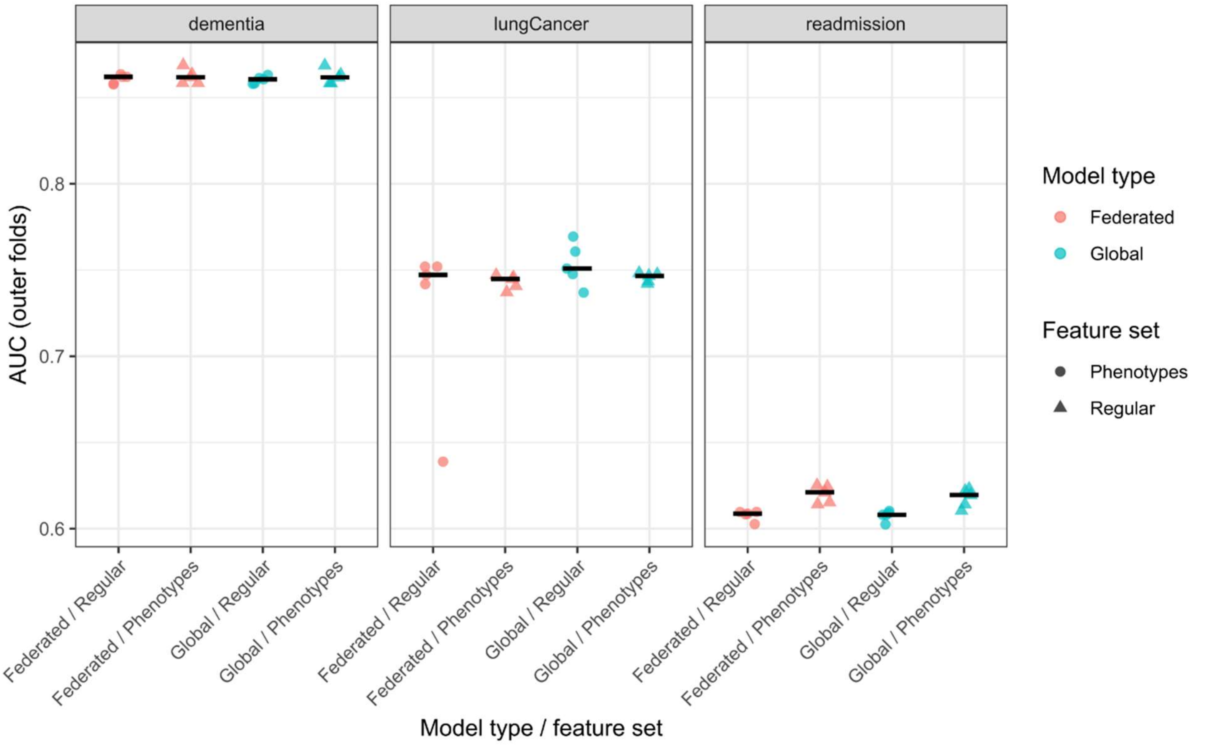
The samples sizes, number of outcomes and features for either regular or phenotype features can be seen in Table 1.

**Table 1:** Sample sizes, outcomes and features per prediction task. For the federated learning the outcomes and sample sizes per client are one fifth of the total.

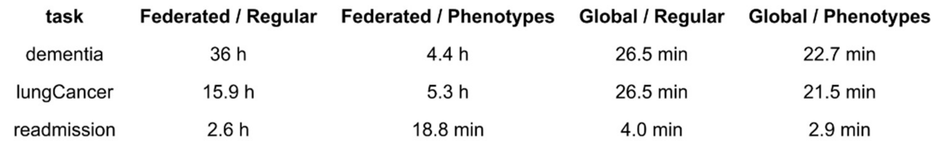
| Prediction task | Total sample size | # Outcome events (%) | Feature types | # Features |
|---|---|---|---|---|
| Dementia | 759120 | 15941 (2.1%) | Regular | 1295 |
| | | | Phenotypes | 62 |
| Lung Cancer | 759189 | 5333 (0.7%) | Regular | 1289 |
| | | | Phenotypes | 63 |
| Readmission | 261316 | 31650 (12.1%) | Regular | 1363 |
| | | | Phenotypes | 63 |

The discrimination performance can be seen in Figure 2a. Across twelve comparisons, the absolute AUC difference never exceeded 0.013 (median 0.002, Figure 2a). Federated training therefore reproduced the discrimination performance of a centrally fitted model irrespective of feature granularity. FDA was 6–83 × slower than pooled training (Figure 2b). The longer runtime stems from its first-order, multi-round optimization and per-round communication over the full parameter vector, whereas the pooled Cyclops solver uses second-order coordinate descent with warm starts and converges in far fewer passes. The gap widens with higher feature dimensionality (Regular > Phenotypes).

a) Nested-CV AUCs per task and model type



b) Runtime per task and model type



| task | Federated / Regular | Federated / Phenotypes | Global / Regular | Global / Phenotypes |
|---|---|---|---|---|
| dementia | 36 h | 4.4 h | 26.5 min | 22.7 min |
| lungCancer | 15.9 h | 5.3 h | 26.5 min | 21.5 min |
| readmission | 2.6 h | 18.8 min | 4.0 min | 2.9 min |

## Conclusion

FDA delivered AUCs indistinguishable from pooled models for three diverse outcomes, confirming that fully automated, data-driven FL can match the gold-standard approach without manually restricting predictors. Switching from Regular to Phenotype features reduced dimension by >95 % (≈63 vs >1300) and cut federated runtime by roughly an order of magnitude while barely affecting performance. Using a restricted set of phenotypes as features is thus a pragmatic strategy to tame the computational overhead of FDA without sacrificing discrimination.

Our FL was simulated by splitting a single database on one host; clients were homogeneous and incurred no network/privacy overhead, which may overestimate agreement with pooled performance and underestimate runtime.

Federated dual averaging achieves near-identical predictive performance to a pooled L1 logistic regression across multiple outcomes and feature representations. Training times are one to two orders of magnitude longer, particularly for higher-dimensional feature spaces. Algorithmic and engineering advances are needed to improve scalability and make fully data-driven FL routinely feasible across distributed healthcare networks.

## References

1. Fridgeirsson EA, Williams R, Rijnbeek P, et al. Comparing penalization methods for linear models on large observational health data. Journal of the American Medical Informatics Association. 2024 Jul 1;31(7):1514–21.

2. John LH, Kim C, Kors JA, et al. Comparison of deep learning and conventional methods for disease onset prediction [Internet]. arXiv; 2024 [cited 2025 Jun 23]. Available from: http://arxiv.org/abs/2410.10505

3. Yuan H, Zaheer M, Reddi S. Federated Composite Optimization. 2021;

4. Reps JM, Wong J, Fridgeirsson EA, et al. Finding a constrained number of predictor phenotypes for multiple outcome prediction. BMJ Health & Care Informatics. 2025;32(1):e101227.

5. Suchard MA, Simpson SE, Zorych I, et al. Massive parallelization of serial inference algorithms for a complex generalized linear model. ACM Transactions on Modeling and Computer Simulation. 2013 Jan 1;23(1).

6. Reps JM, Schuemie MJ, Suchard MA, et al. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association. 2018 Aug 1;25(8):969–75.