frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 29 February 2024
DOI 10.3389/fdata.2024.1266031

OPEN ACCESS

EDITED BY
Tuan D. Pham,
Queen Mary University of London, United Kingdom

REVIEWED BY
Pietro Pinoli,
Polytechnic University of Milan, Italy
Wenan Chen,
Mayo Clinic, United States

*CORRESPONDENCE
Dmitry Kolobkov
✉ dmitry.s.kolobkov@gmail.com

†These authors share first authorship

RECEIVED 26 July 2023
ACCEPTED 31 January 2024
PUBLISHED 29 February 2024

CITATION
Kolobkov D, Mishra Sharma S, Medvedev A, Lebedev M, Kosaretskiy E and Vakhitov R (2024) Efficacy of federated learning on genomic data: a study on the UK Biobank and the 1000 Genomes Project.
Front. Big Data 7:1266031.
doi: 10.3389/fdata.2024.1266031

Efficacy of federated learning on genomic data: a study on the UK Biobank and the 1000 Genomes Project

Dmitry Kolobkov[1,2*†], Satyarth Mishra Sharma[1,3†], Aleksandr Medvedev[1,3], Mikhail Lebedev[1], Egor Kosaretskiy[1] and Ruslan Vakhitov[1]

[1]GENXT, Hinxton, United Kingdom, [2]Laboratory of Ecological Genetics, Vavilov Institute of General Genetics, Moscow, Russia, [3]Center for Artificial Intelligence Technology, Skolkovo Institute of Science and Technology, Moscow, Russia

Combining training data from multiple sources increases sample size and reduces confounding, leading to more accurate and less biased machine learning models. In healthcare, however, direct pooling of data is often not allowed by data custodians who are accountable for minimizing the exposure of sensitive information. Federated learning offers a promising solution to this problem by training a model in a decentralized manner thus reducing the risks of data leakage. Although there is increasing utilization of federated learning on clinical data, its efficacy on individual-level genomic data has not been studied. This study lays the groundwork for the adoption of federated learning for genomic data by investigating its applicability in two scenarios: phenotype prediction on the UK Biobank data and ancestry prediction on the 1000 Genomes Project data. We show that federated models trained on data split into independent nodes achieve performance close to centralized models, even in the presence of significant inter-node heterogeneity. Additionally, we investigate how federated model accuracy is affected by communication frequency and suggest approaches to reduce computational complexity or communication costs.

KEYWORDS

federated learning (FL), phenotype prediction, ancestry prediction, machine learning, data collaboration, genomics, polygenic scores

1 Introduction

Here, we describe the current trends and policies concerning genomic data, its usage for phenotype and ancestry prediction, as well as current practices of federated learning and privacy-enhancing mechanisms.

1.1 Availability of genomic data

The last decade has seen a rapid increase in the amount of genomic data due to the improvement of sequencing technologies and the promise of big data studies in healthcare. With genotyping costs going down, the pool of genomic data also becomes less centralized as more organizations, both commercial,

such as genetic testing companies, and non-profit, such as biobanks, accumulate vast collections of genomes. This decentralization, coupled with data-hungry genome-wide machine learning approaches, raises a need for data collaboration. However, access to genomic data is usually restricted due to its sensitive nature and the harmful consequences of possible data leakage, such as deanonymization and genetic discrimination (Joly et al., 2017; Bonomi et al., 2020; Chapman et al., 2020).

Data holders may share aggregated data such as summary statistics for genome-wide association studies (GWAS) to allow easy access with a reduced risk of exposing sensitive information. The summary statistics can be analyzed jointly via meta-analysis (Evangelou and Ioannidis, 2013; Ray and Boehnke, 2018). However, summarizing involves a loss of information which affects model performance.

Due to new governmental policies, data custodians of large healthcare cohorts started to store sensitive individual-level data in secure data havens, which are accessible via trusted research environments (UK Health Data Research Alliance and NHSX, 2021; Kavianpour et al., 2022; Mayo et al., 2023). Federations of such protected data silos are being established and new federated analysis frameworks are being developed, which allow joint analyses of data from multiple data silos (The Global Alliance for Genomics and Health, 2016).

1.2 Phenotype prediction

Phenotype-from-genotype prediction aims to score an individual's genetic liability to a certain phenotype, usually, a disease, which can identify risk groups and assist diagnostics (Lewis and Vassos, 2020). To build a predictive model, one has to obtain either individual-level data where each sample has two alleles for each included genetic variant (SNP) or summary-level data where each SNP has an allele frequency.

Models trained on individual-level data typically yield higher predictive performance as they learn the joint SNP distribution. However, this sensitive data can typically be accessed only with an approved research application.

On the other hand, summary-based models, or polygenic scores, are trained on publicly available GWAS-derived summary statistics and can even incorporate outputs of multiple GWAS using meta-analysis. However, polygenic score models are typically based on assumptions that reduce their applicability to samples with ancestry different from the ancestry of the training set. For instance, multiple studies show poor portability of polygenic scores to other ancestry groups (Gurdasani et al., 2019; Martin et al., 2019a; Privé et al., 2022). This is caused by inter-population differences in allele frequency (Durvasula and Lohmueller, 2021) and variant effect size (Shi et al., 2021), as well as different linkage disequilibrium patterns (Amariuta et al., 2020).

In this paper, we consider only individual-level data, since summary-level data is typically publicly available and does not require federated learning to keep it private.

1.3 Ancestry prediction

Genetic ancestry prediction from SNPs has two common uses. First, it is a product that genetic testing companies provide to their customers (Kirkpatrick and Rashkin, 2017). Despite its not purely scientific purpose, predicted ancestry is an important factor in attracting new customers to provide their DNA samples and, thus, increases the amount of available individual-level genetic data. Second, due to poor polygenic score cross-ancestry portability, per-ancestry summary statistics (Buniello et al., 2019) and polygenic score (Lambert et al., 2021) catalogs have been established. As self-reported ancestry is often noisy, ancestry prediction is a promising tool to be used in phenotype prediction and pharmacogenomics (Yang et al., 2021).

The task of predicting ancestry is closely related to inferring genetic population structure. As ancestry differences comprise the major part of human genetic variation, population structure is well-described by top eigenvectors of the covariance matrix obtained from genetic variants. As a consequence, these principal components are commonly included in phenotype prediction models as covariates to control for population structure (Price et al., 2006). A common approach to estimate the ancestry of unlabeled samples is to project it onto the principal component space of a labeled reference panel (Privé et al., 2020), such as the 1000 Genomes Project (The 1000 Genomes Project Consortium, 2015).

1.4 Federated learning

Federated learning involves training a model locally on clients (data nodes) and sending parameter updates to a server (central hub) where these updates are aggregated into a new set of model parameters which are then sent back to the clients in the next round of training, also called a communication round (Yang et al., 2019). Unlike conventional "centralized" machine learning, federated learning does not require assembling data at a single location which saves communication costs and, more importantly, enhances data security since the client's data is not disclosed to the server nor to other clients.

Federated learning features different strategies, which vary in aggregation methods on a server, in the training process on clients and in the communication frequency between a client and a server. Different strategies may be preferable in different data distribution scenarios: (i) cross-device (many clients with little data) or cross-silo (few clients with a lot of data); (ii) varying degrees of inter-client heterogeneity (dissimilarity); (iii) unavailable or straggling clients. This study considers the cross-silo scenario which is the most typical for genomic data where clients, such as hospitals, biobanks, and genetic testing companies may be dissimilar due to differing genetic populations.

1.5 Privacy issues of federated learning

The key privacy-preserving mechanism in federated learning is keeping the data at data silos where it is stored and sending model parameters, instead of the data, to and from data silos. While

sending the model parameters is more secure than transmitting the sensitive data, this alone may not be enough to guarantee data privacy, since the model parameters may leak information about the data, for instance through gradient updates during the training process (Lim and Chan, 2021; Mothukuri et al., 2021). It has been shown that phenotype prediction in a federated setting can be vulnerable to deanonymization attacks (Wjst, 2010). Thus, additional privacy enhancing mechanisms may be required. These include differential privacy (Abadi et al., 2016), secure multiparty computation (Damgård et al., 2012) and trusted execution environments (Mo et al., 2021), see Bonomi et al. (2020) for a comprehensive survey of privacy-preserving mechanisms.

1.6 Federated learning for genomics and healthcare

Trusted research environments forbid moving sensitive data out of data havens, which prevents the downloading and combining of data from multiple sources to train conventional machine learning models. Federated learning, on the other hand, is designed to operate on isolated data silos. Therefore, there is a growing demand for federated learning infrastructures for healthcare data (Nik-Zainal et al., 2022; Alvarellos et al., 2023). Researchers have applied federated learning to a variety of healthcare data, including electronic health records (Vaid et al., 2021), medical images (Li et al., 2020) and wearables (Chen et al., 2020). A number of surveys describe the applications, prospects, challenges and privacy concerns of federated learning in healthcare (Rieke et al., 2020; Xu et al., 2021; Joshi et al., 2022). A number of privacy-preserving techniques for genomic data have been proposed, such as federated GWAS (Sadat et al., 2018; Nasirigerdeh et al., 2020) and federated PCA for GWAS (Hartebrodt et al., 2021). However, to the best of our knowledge, the efficacy of federated learning for predictions from full-scale genomic data has not been extensively investigated. Training a federated model on genomic arrays poses additional challenges typical to omics data, such as the vast number of non-independent features (genetic variants).

1.7 Outline and scope of the paper

The paper is structured as follows. In the Results, we first compare federated, local and centralized models trained to predict eight phenotypes from SNP data of the UK Biobank. In our second experiment, we analyze the behavior of federated models for ancestry-from-genotype prediction on the highly heterogeneous 1000 Genomes data. Finally, we provide recommendations on a training schedule by varying the number of communication rounds and local epochs in each round depending on the system bottleneck to achieve high accuracy. In the Section 3, we detail the data preprocessing and model training for both experiments. In the Section 4, we overview and analyze the results and suggest directions for future research.

In both experiments, we use well-established and accurate models and train them using the standard FedAvg strategy on artificially-separated datasets to provide a baseline and focus on the analysis of federated model behavior. Here, we do not consider additional privacy-enhancing mechanisms. Future studies may focus on using FL with more advanced models, designing novel FL strategies, protecting federated models against the attacks, as well as train federated models on multiple real datasets, such as biobanks.

Our main contributions are as follows. First, we show the efficacy of federated models on genomic data. We demonstrate that federated models can be successfully trained to achieve almost the same performance as centralized models, which are not limited by privacy restrictions, and by a large margin outperform local models, trained in compliance with privacy restrictions. Second, a clear issue for federated learning on large genomic datasets is that it may be jeopardized by high computational load and large amount of communication, the latter of which also increases system vulnerability to attacks. We demonstrate how, depending on the system bottleneck, to choose the number of local epochs, frequency of communication between the nodes and the quality of federated data preprocessing in order to keep performance high in the presence of high inter-node heterogeneity. In summary, we provide fundamental evidence of a successful use of federated learning on genomic data which, we hope, will encourage genomic collaboration and future research.

2 Results

In this section, we compare federated, centralized and local phenotype-from-genotype prediction models on the UK Biobank dataset. Further, we analyze the behavior of federated models in the presence of high data heterogeneity on the 1000 Genomes dataset and investigate how the performance of federated models depends on the amount of client-server communication and the number of epochs in a communication round. Table 1 provides a high-level overview of the two experiments.

2.1 Phenotype prediction from UK Biobank data

In this experiment, we mimic the situation where genomic data is stored in multiple large silos, such as hospitals, within the same country. We split the UK Biobank data into 19 datasets according to sample collection centers in different parts of the UK. Some inter-node heterogeneity is present due to the correlation between the UK's genetic population structure and geography (Agrawal et al., 2020). After a standard quality control (QC), we reduced dimensionality by conducting GWAS on each node and selecting top SNPs. Then, we trained local, federated and centralized Lassonet neural networks (see Section 3) of identical architecture with selected SNPs, sex and age as features. Advanced lasso-based models are considered to be state-of-the-art for phenotype prediction from individual-level genomic data (Prive et al., 2018; Qian et al., 2020). Here, we chose a basic Lasso-based model to focus on the relative performance of federated, centralized and local models. The experiment setup is visualized in Figure 1 and described in more detail in the Section 3.

Figure 2 displays R^2 performance of six out of 19 local models, federated (FedAvg, 8 epochs in a communication round, see Section

TABLE 1 Overview of the two experiments.

Prediction	Phenotype-from-genotype	Ancestry-from-genotype
Dataset	UK Biobank	1000 Genomes
Population	Mostly White British	5 superpopulations
Use case for	Collaboration within a country	Cross-continental collaboration
Node heterogeneity	Low	High
Client nodes	19	5
Node sizes	12k–42k samples	509–682 samples
Comparison	Federated vs. centralized vs. local models	Different federated models
Predictive task	Regression	Multiclass classification
Input features	10k GWAS-selected SNPs + age + sex	20 PCs
Prediction target	Quantitative phenotype	26 population classes
Model description	LASSO	Fully-connected MLP
Loss function	Mean Squared Error + l1 penalty	Cross-entropy
Metric	R^2 (coefficient of determination)	Accuracy
Validation	10-fold cross-validation: eight folds - train, one fold - validation, one fold - test	10-fold cross-validation: eight folds - train, one fold - validation, one fold - test

3) and centralized models on the same test set, for eight continuous phenotypes. The six displayed nodes were selected, prior to model training, to demonstrate the whole range of sample sizes. The considered phenotypes were selected based on previous heritability estimates (Sinnott-Armstrong et al., 2021). On each node two local models were trained: one trained on top SNPs from local GWAS and one trained on top SNPs from meta-GWAS (see Section 3). The federated models used SNPs from meta-GWAS and the centralized models used SNPs from centralized GWAS. Local and centralized covariates-only (sex and age) models were also trained as a baseline.

We trained local and centralized models with (i) "native" features, i.e., SNPs derived from local and centralized GWAS, to represent end-to-end solutions, and (ii) SNPs yielded by meta-GWAS so that local, centralized and federated models can be compared on exactly the same set of features. For centralized models, performance on centralized GWAS SNPs and meta-GWAS SNPs was very similar for all phenotypes, thus, only the former was included in the Figure 2. Local models tend to perform better on meta-GWAS SNPs compared to local GWAS SNPs, most likely because SNP selection via a local GWAS tends to yield more false positives due to an insufficient number of samples. For local models, we see a natural trend that performance improves as the node size grows. Federated models outperform all local models and get close to centralized models.

2.2 Ancestry prediction from 1000 Genomes data

Despite its lower clinical significance than predicting complex traits from genotype, we now consider ancestry-from-genotype prediction models. They can be trained using a smaller number of samples and are lighter and less computationally expensive, which allows us to conduct extensive simulations to investigate federated learning in more detail. Here, we mimic the situation where genetic testing companies from different parts of the world collaboratively predict ancestry and split the 1000 Genomes Project data into 5 isolated nodes based on sample superpopulation (African, Native American, East Asian, European, Southern Asian), thus getting high inter-node heterogeneity. After a standard QC, we reduced dimensionality by applying federated PCA (Hartebrodt and Röttger, 2022) to pruned SNPs and then trained local, federated and centralized multilayer perceptrons (MLPs) of identical architecture. The experiment setup is visualized in Figure 3 and described in more detail in the Section 3.

Our goal here is to investigate the performance of federated models as a function of communication between the clients (nodes) and the server in the presence of significant cross-client heterogeneity. For this, we compare FedAvg strategies with a different number of epochs in a round of communication. The training process of federated models is displayed in Figure 4.

Figure 4A compares the validation loss of the centralized and federated models. Federated models show a clear trend that the more communication between the server and the clients, i.e., the more rounds and the fewer local epochs in each round, the faster the convergence is. For the centralized model, convergence was fast, however, it started from a higher loss value because of the random class assignment in a multiclass classification initialization: a centralized model solves a problem for 26 classes whereas each of the five superpopulation nodes has fewer classes (26 in total).

Figure 4B shows the evolution of the client training loss of the FedAvg model with 32 local epochs in a round of communication. Here, the peaks correspond to the initial evaluation of the aggregated parameters sent from the server to the client. Parameter aggregation on the server results in an increase of the local loss as the aggregated parameters are a weighted average of parameters optimized on different data distributions (due to inter-node heterogeneity), then the loss starts decreasing as the model starts fitting to the local data.

2.3 Practical considerations of server-client communication

Figure 5 shows the accuracy of federated models as functions of the number of total epochs (computational complexity) and rounds (communication). In compliance with Figures 4A, 5A shows that for heterogeneous data, increasing communication between the clients and the server leads to higher accuracy. On the other hand, Figure 5B shows that for a limited number of rounds, it is beneficial to train locally for a larger number of epochs. Thus, depending

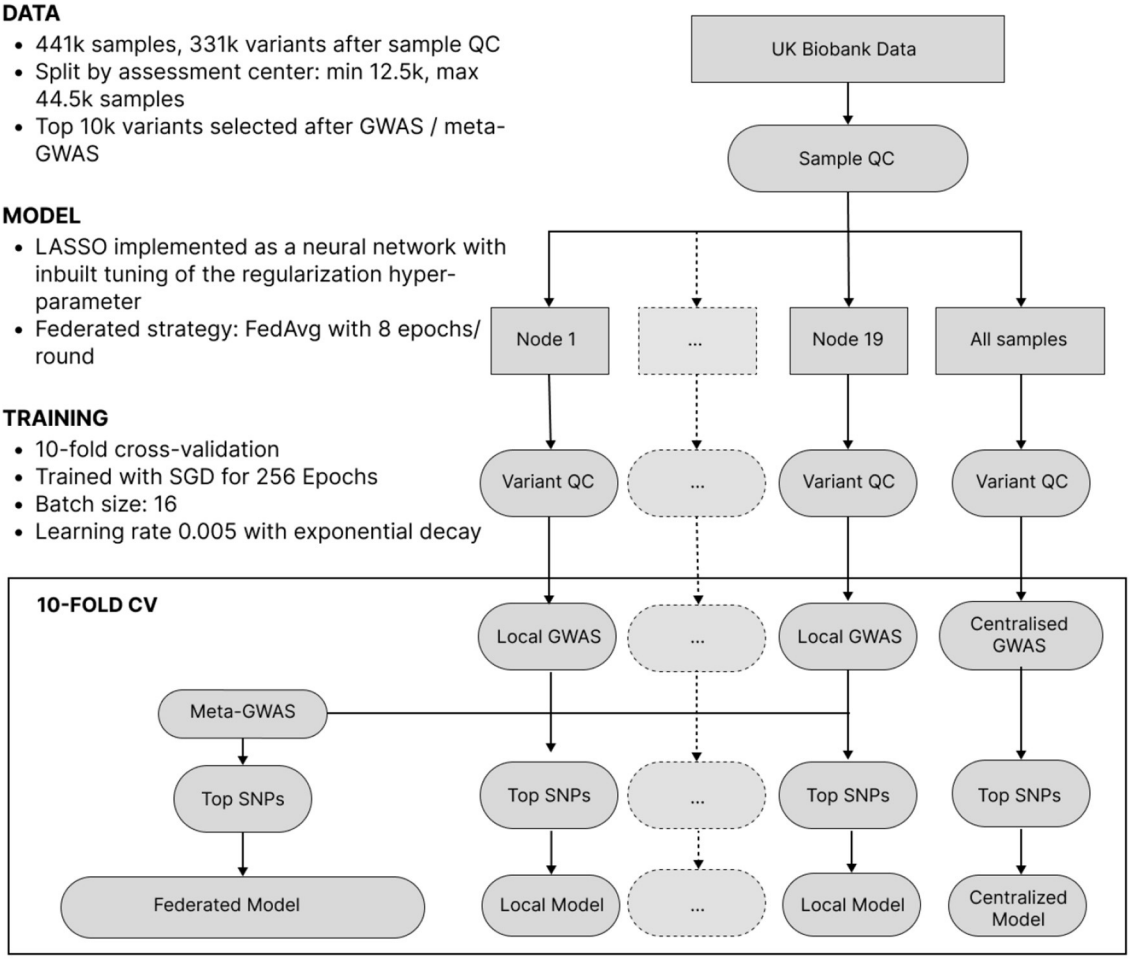

FIGURE 1
The workflow of phenotype prediction from UKB data. Data is split into nodes and undergoes per-node variant QC. Then, within a cross-validation loop dimensionality is reduced by selecting the most significant SNPs via a local or meta-GWAS. Finally, models are trained.

on what is the bottleneck of the system, communication or computational complexity, different federated learning strategies may be preferable.

A fully federated solution requires all data to be prepared in a federated manner as well. In the case of ancestry prediction, dimensionality reduction is usually conducted via PCA, thus, we first pruned SNPs as displayed in Figure 3 to decrease computational load and then utilized federated PCA using the P-stack algorithm as described in Hartebrodt and Röttger (2022). The amount of communication used in federated PCA linearly depends on the number of input SNPs which affects the model accuracy. Hence, if communication is limited, one can spend more on the PCA step by including more SNPs or use more communication rounds for model training. Figure 6 shows our rationale behind choosing the pruning parameters that determine the number of input SNPs for federated PCA. We also validate the federated PCA approach and show that the centralized classifier performs similarly on federated and centralized PCs.

3 Methods

In this section, we describe the federated learning strategies used in this study and then detail the data preprocessing and model training and evaluation for both experiments.

3.1 Federated learning strategies

This study aims to assess the applicability of federated learning to genomic data. In this paper, we consider only "global" models that aim to perform well on all varieties of data split between isolated datasets. A survey of personalization methods for federated models is provided in Kulkarni et al. (2020).

Intuitively, we expect a good federated model to perform considerably better than the local models, trained on data in a single node, and slightly worse than a centralized model that trains on all of the data together. In this case, a federated model delivers all

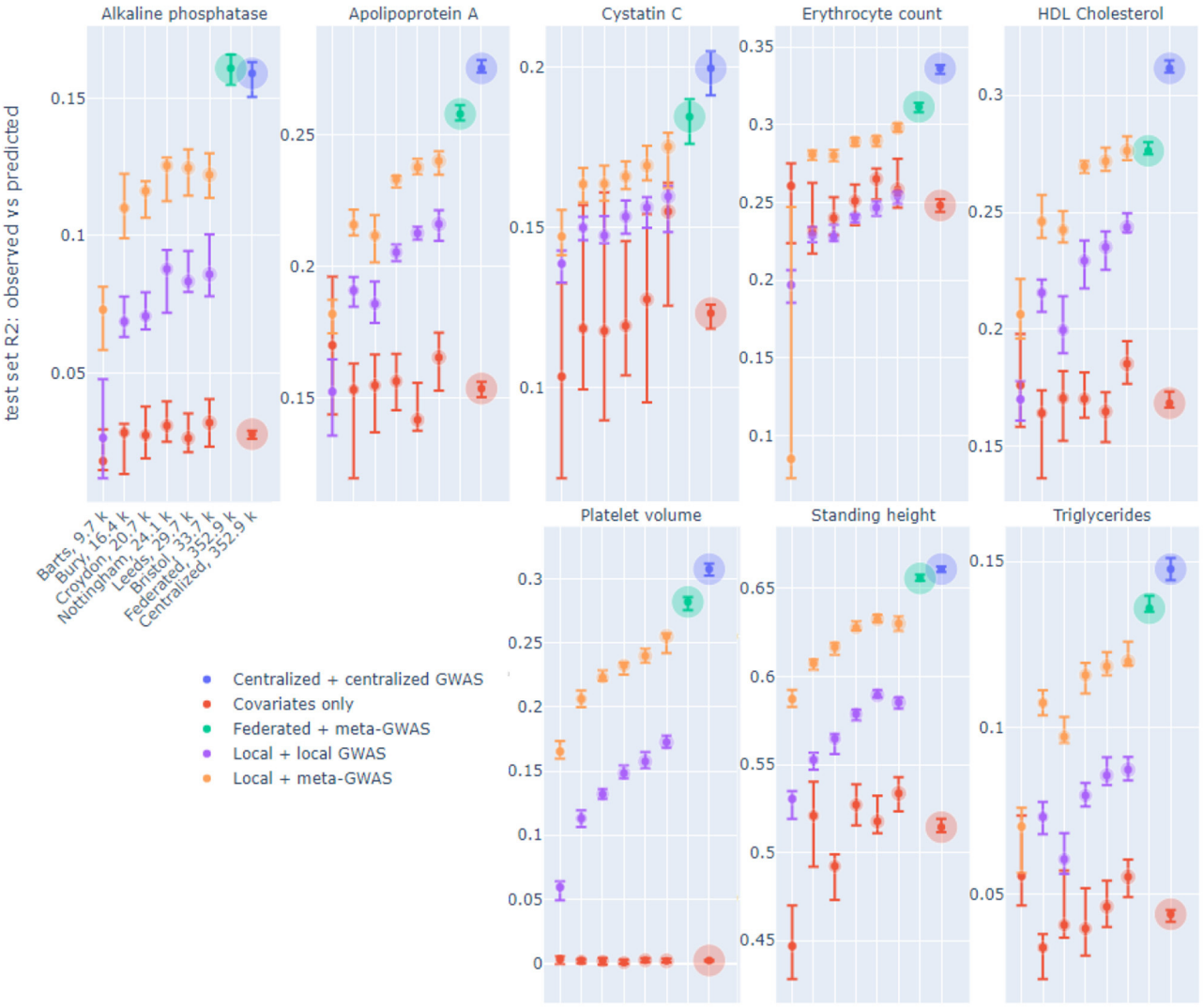

FIGURE 2
R^2 of observed vs. predicted phenotype of local, federated and centralized models evaluated on the same test set on UK Biobank data split by assessment center. Facets correspond to the phenotypes. The x-axis shows six of 19 local nodes, selected to represent a range of node sizes, in increasing size order, which is the same for each phenotype. The number of training samples in each node is indicated in the x-axis labels, which bubble size also corresponds to. Color corresponds to the model type. Each dot corresponds to the median across 10-fold cross-validation for each performance metric, and the intervals shown are 0.1–0.9 quantiles obtained by dropping the largest and smallest of 10 values.

benefits of federated learning at the cost of a small reduction in performance compared to a centralized model.

Here, we consider the FedAvg strategy (McMahan et al., 2017) with a different number of epochs in a communication round. In our case, where the number of clients is low, the pseudocode is displayed in Algorithm 1.

In the presence of significant inter-node heterogeneity, i.e., when the local data distribution on the clients is different from the global distribution, the global model tends to overfit to local data causing "client drift" that slows or prevents convergence (Li et al., 2019). Client drift can be decreased by limiting the amount of training on a client in a single round, i.e., increasing communication between a client and a server. In this paper, we utilize FedAvg with different amounts of server-client communication by varying the number of communication rounds K but keeping the total number of local epochs KE constant.

3.2 The UK Biobank dataset and data processing

An overview of our data processing pipeline for experiments on the UK Biobank dataset can be seen in Figure 1. All samples underwent quality control using PLINK (Chang et al., 2015) to remove individuals with an insufficient number of genotyped variants (6% missingness cutoff) and related samples with a KING (Manichaikul et al., 2010) cutoff 0.0884 corresponding to second-degree relatives. For local and federated but not centralized models, the data was split into 19 datasets according to the data collection center (UKB data-field 54). Data collection centers with less than 10k samples were excluded. The size of the datasets after QC ranged from 12.1k (Barts) to 42.1k (Bristol), for a total of 441k samples. See Supplementary Table 1 for the exact breakdown of node sizes. For each dataset separately, variant QC was conducted by filtering

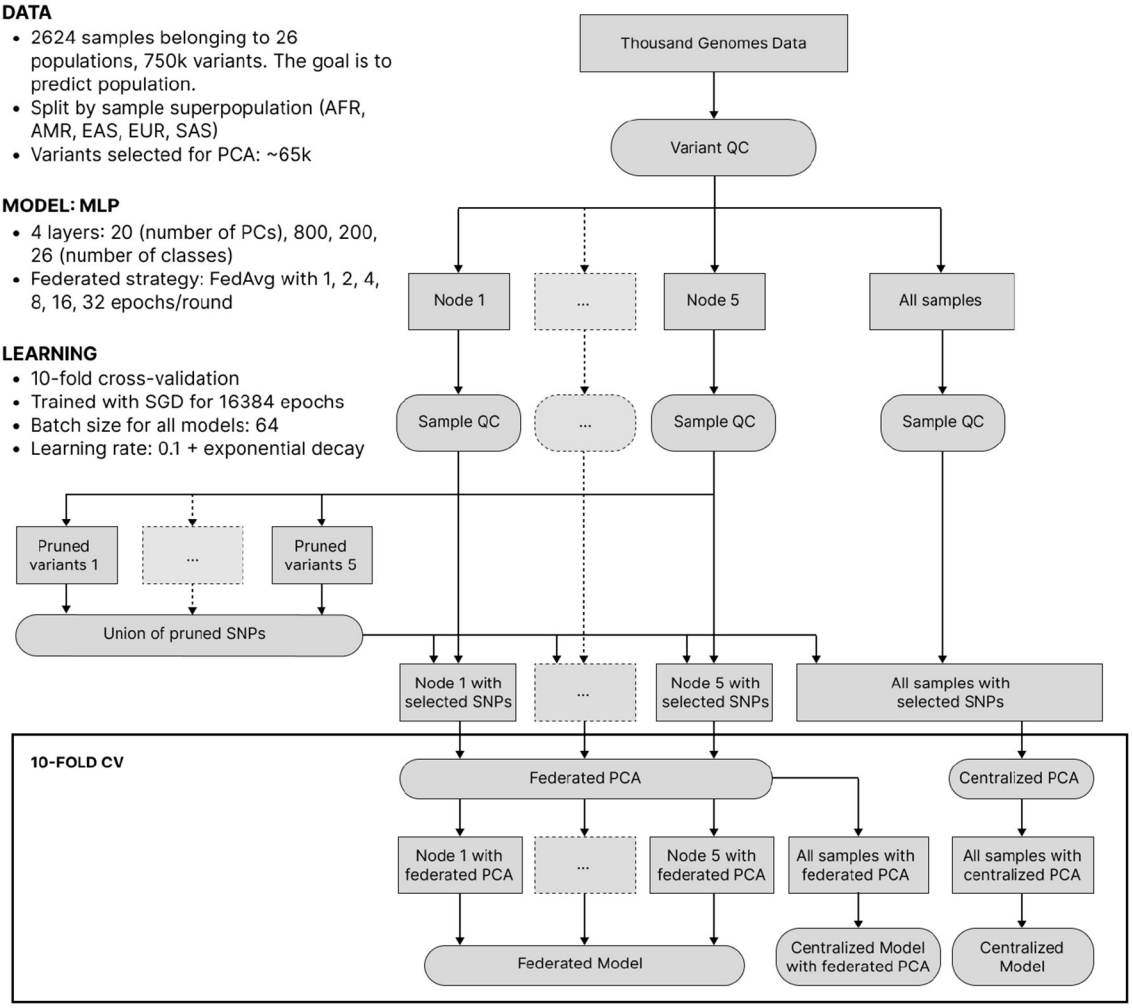

FIGURE 3
The workflow of ancestry prediction from 1,000 Genomes data. Data was split into nodes according to sample superpopulation. The union of variants pruned on each node was taken for all nodes to have the same features. Federated/centralized PCA was used to further decrease dimensionality. Finally, federated and centralized models were trained.

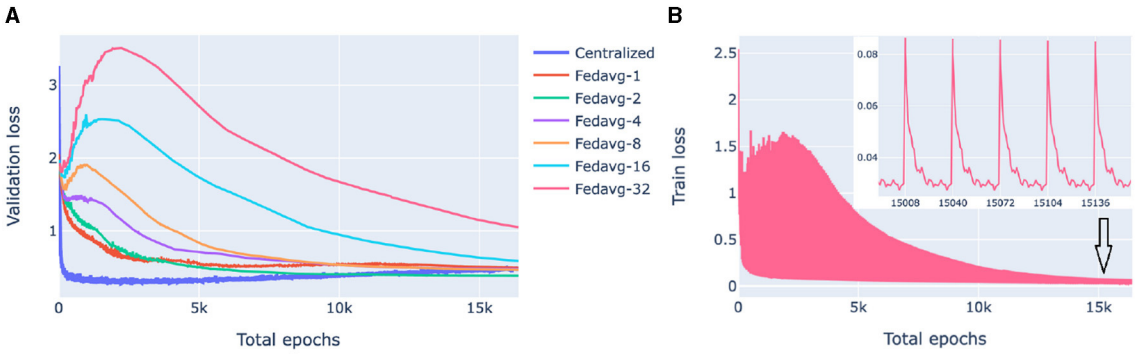

FIGURE 4
Loss behavior of federated and centralized models. FedAvg strategy with 1, 2, 4, 8, 16, 32 epochs/round was used. **(A)** Validation loss of federated models with a different number of communication rounds. For each model and epoch, a median loss over 10-fold cross-validation is shown. **(B)** The training process of a 32-epochs-in-round model on a client. Every 32nd epoch contains two points: one before the start of the epoch when the client receives parameters from the server and one after the first epoch of a round.

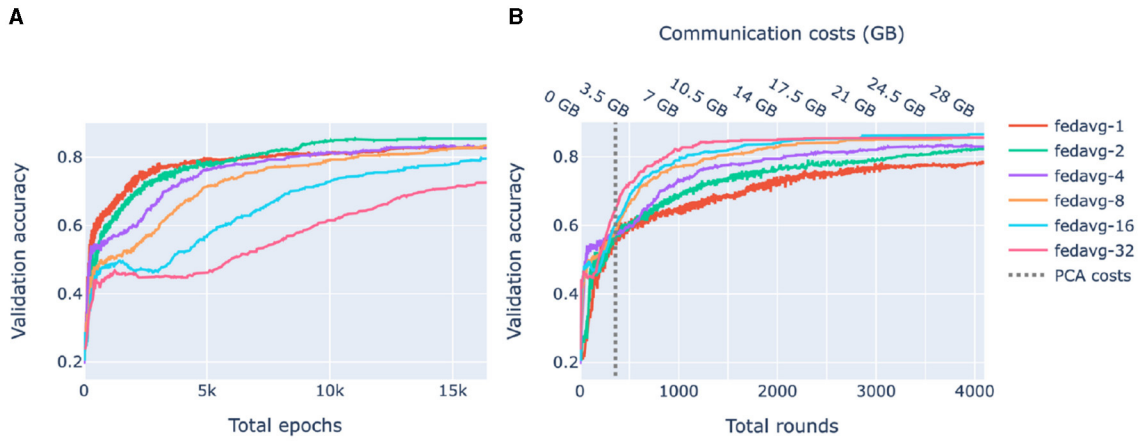

FIGURE 5
Validation accuracy as a function of complexity and communication. FedAvg strategy with 1, 2, 4, 8, 16, 32 epochs/round was used. Each shown value is a median over 10-fold cross-validation. **(A)** Accuracy of federated models as a function of the total number of epochs. **(B)** Accuracy of federated models as a function of the amount of communication between the server and the clients. The dashed line corresponds to the amount of communication used by federated PCA.

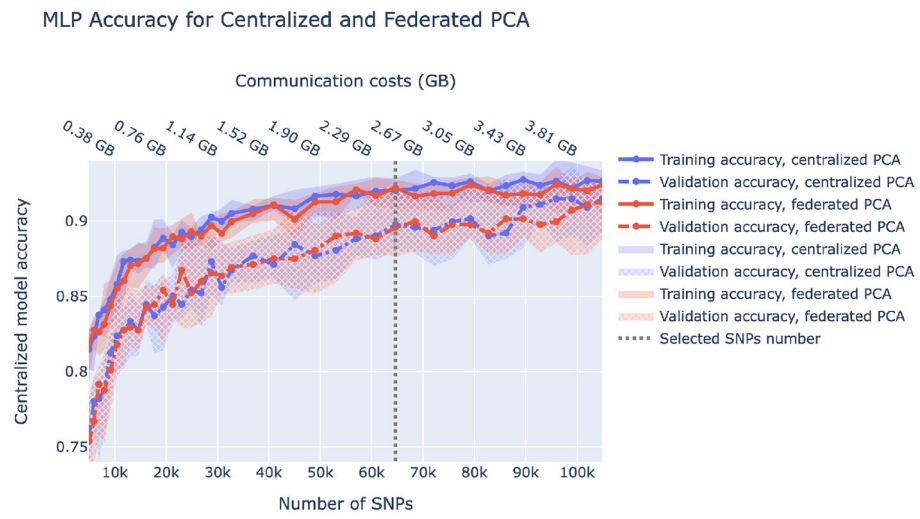

FIGURE 6
Centralized model accuracy as a function of the number of SNPs used for dimensionality reduction via centralized and federated PCA. Solid/dashed line corresponds to train/validation model accuracy, and blue/red corresponds to centralized/federated PCA. Shaded areas correspond to 0.1–0.9 quantiles based on 10-fold cross-validation. Vertical dashed lines correspond to the number of SNPs we chose to be used in downstream analysis.

out variants by rarity (5% minor allele fraction cutoff), missingness (2% cutoff), and a Hardy-Weinberg equilibrium p-value threshold of 10e-6. Each dataset was split into 10 folds for cross-validation where at each time eight folds are used as training data, one as validation data (to be used for regularization parameter selection), and one as test data. Next, we reduced dimensionality by selecting the 10,000 most significant SNPs with a GWAS conducted on the training data using age, sex, and 20 principal components from SNPs as covariates. These 10,000 SNPs were combined with age and sex as the input features for our predictive model. For experiments with federated models, feature selection was done by performing a random-effects meta-analysis, which we chose over a fixed-effect analysis due to better predictive performance on the 9 phenotypes we considered. The meta-analysis was performed using PLINK by aggregating information from the GWAS reports of the individual datasets and then selecting the top 10,000 SNPs.

3.3 Phenotype prediction model

We implemented a LASSO model for phenotype prediction. LASSO is a regularized linear model, with the MSE loss plus l1 penalty (Tibshirani, 1996). LASSO excels in high-dimensional problems, such as phenotype prediction from multiple SNPs, because LASSO loss is convex and has a built-in feature selection (Hastie et al., 2015), which makes it fast to optimize and more interpretable. LASSO models are commonly used on large genomic

```
procedure SERVER-SIDE
    Initialize ω ← ω_0                          ▷ Initialize model weights
    for r ∈ 1, .., R do
        for k ∈ 1, .., K do <in parallel>
            ω_r^k ← Client-side(k, ω)           ▷ Parallel local optimization
        end for
        ω ← Σ_{k=1}^{K} (n_k / n) ω_r^k         ▷ Weighted average of client weights
    end for
end procedure
procedure CLIENT-SIDE(k, ω)
    for i ∈ 1, .., E do
        for batch b do
            ω ← ω − η∇F(ω, b)                   ▷ Batch gradient descent
        end for
    end for
    return ω
end procedure
```

Algorithm 1. FedAvg for a small number of clients. K—number of clients, n_k—number of samples on kth client, R—number of communication rounds, E—number of local epochs in a round, η—local learning rate.

datasets and are solved iteratively to save memory consumption (Prive et al., 2018; Lello et al., 2019; Qian et al., 2020). A LASSO problem can be solved using coordinate descent (Lello et al., 2019) or gradient descent. We chose to solve LASSO using gradient descent because it can be easily implemented on top of existing deep learning and federated learning frameworks, such as PyTorch (Paszke et al., 2019) and Flower (Beutel et al., 2020).

However, with gradient descent, the LASSO problem can only be solved for a single value of the regularization parameter λ. Since λ determines model performance and generalization ability, one typically trains multiple models with different values of λ and selects the one with the best validation metric. We implemented this procedure as a linear neural network which efficiently trains LASSO models with a range of λ values in parallel and offers built-in model selection. We call this implementation Lassonet.

For each cross-validation fold, all local models were trained and validated on a single corresponding node, federated models were trained and validated on all nodes but in isolation, whereas centralized models were trained and validated on train and validation sets from all nodes merged together. For each cross-validation fold, all models were tested on exactly the same "centralized" test set that consisted of test sets from all nodes merged together.

Centralized and local Lassonet models were trained using PyTorch and PyTorch Lightning (Falcon et al., 2020) on the Zhores cluster node with Nvidia V100 GPU with 16 GB VRAM and up to 160 GB of RAM (Zacharov et al., 2019). We used the SGD optimizer with learning rate 5e-3, learning rate decay 0.99, batch size 16 and trained Lassonet for 256 local epochs for each run on the UK Biobank data. We implemented federated models using the Flower framework. Here, we aggregated validation loss from models with the same λ across clients each round, and then chose the model with the best validation loss to be evaluated on the test set.

3.4 The 1000 Genomes Project dataset and data processing

The 1000 Genomes array contains about 750 thousand genetic variants (SNPs) of 2,624 samples of 26 genetic populations belonging to five superpopulations of East Asians (EAS), Southern Asians (SAS), Europeans (EUR), Africans (AFR), and Native Americans (AMR).

Our data processing workflow sketched in Figure 3 was performed in the following order. First, we conducted variant QC in PLINK keeping genetic variants with minor allele frequency >5% and missing call rates <2%. Next, we split the samples into five isolated nodes according to sample superpopulations. See Supplementary Table 2 for the exact breakdown of node sizes. Then, we conducted sample QC on each node separately in PLINK keeping non-related (KING relatedness cutoff 0.0884 that corresponded to second-degree relatives) samples with missing call rates <6%.

Next, we reduced dimensionality by first pruning variants on each node separately in PLINK, then taking a union of the remaining variants across the nodes to get a single variant set (only variant IDs are communicated between nodes) yielding about 65 thousand SNPs. Then, data on each node was split into 10 folds for cross-validation where at each time eight folds are used as training data, one as validation data (to be used for early stopping during model training) and one as test data. Finally, we conducted federated PCA and, alternatively, centralized PCA on the training set and extracted the top 20 principal components. To reduce the dimensionality of validation and test sets, we projected them onto the training PC space. The influence of the pruning strictness and the federated vs. centralized PCA is displayed in Figure 6.

3.5 Federated PCA for dimensionality reduction

The standard way to reduce dimensionality for ancestry-from-genotype prediction is by using the principal component transformation as it is well-known that PCs of genetic variants retain genetic population structure in the dataset (Patterson et al., 2006). The federated models we used in this study require client datasets to have the same feature space, therefore the PCs have to be obtained collaboratively. Since computing PCs centrally discloses the data and therefore compromises the purpose of downstream federated learning, we employed the federated PCA approach.

We utilized the P-STACK method as described in Hartebrodt and Röttger (2022), which involved sending a local eigenvalues vector and an eigenvector matrix from each client to the server. Further, the server stacks local PCA components and then performs a singular value decomposition (SVD) of the obtained joint matrix. To perform an exact PCA, we used the maximum available number of eigenvectors on each client, which equals the number of client samples minus one. When the number of PCs is fixed, the size of the eigenvector matrix depends only on the number of genetic variants. We used PLINK to prune genetic variants, i.e., removed variants in close linkage disequilibrium. Pruning was conducted on each individual node and then a union of remaining on each

node variants was taken. This allowed us to reduce communication costs for federated PCA and significantly shrink RAM consumption while running SVD on the server. When SVD is completed, the resulting eigenvector matrix is sent back to the clients after which each client is able to perform the PC transformation into the joint feature space.

3.6 Ancestry prediction model

Due to the fact that cross-populational differences are captured well by top principal components, any reasonably good model trained on the first 20 PCs will predict ancestry accurately. Since the goal of our experiment is to analyze the behavior of a generic federated model, we trained a standard multi-layer perceptron neural network on these 20 federated PC features.

We used a fully-connected neural network with two hidden layers of size 800 and 200, respectively, 20 input and 26 output neurons, the total of 182K parameters, and the *selu* activation function. The model outputted raw scores of a sample belonging to the each of 26 populations. The cross-entropy loss was used, which is standard for multiclass classification due to its properties of being smooth and convex, making it fast to optimize with gradient descent (Jung, 2022). Similarly to the phenotype prediction experiment, 10-fold cross-validation was used, with the train set consisting of eight folds and the validation and test sets consisting of one fold each. We trained the model for 16384 local epochs with batch size 64, learning rate 0.1 and exponential learning rate decay with γ 0.9999. We trained it on CPU-only machines with 4 CPUs and 8–16 GB of RAM.

4 Discussion

The promise of federated learning for healthcare, and genomics in particular, is a result of two powerful trends. First, machine learning models require a lot of data to train and their applicability depends on the diversity of the training dataset. Training the model on diverse data obtained from multiple sources reduces confounding by population genetics, experimental design, etc. and generally improves performance on external data. Second, the growing awareness of the sensitivity of healthcare data and the harmful consequences of its leakage encourages data custodians to restrict data access, e.g., by requiring an application approval and then granting access only within a trusted research environment (Mansouri-Benssassi et al., 2021; Kavianpour et al., 2022). This makes merging multiple datasets at a single location challenging, thus discouraging training conventional centralized models.

Nevertheless, the applicability of federated learning to individual-level genomic data has not been studied extensively. In this paper, we analyzed the behavior of federated models in two scenarios: phenotype-from-genotype prediction on the UK Biobank data and ancestry-from-genotype prediction on the 1000 Genomes Project data. We first showed that federated models are almost as accurate as centralized models and considerably more accurate than local models for predicting multiple phenotypes from genomic data.

It would have been interesting to split UKB to nodes by ethnic background and see how FL performs in the presence of higher node heterogeneity. However, non-European nodes in UKB would have less than 10000 samples which may not be enough to make robust predictions of complex phenotypes. On the other hand, even hundreds of samples are enough to make accurate ancestry predictions, due to the fact that ancestry has much more genetic variation than complex phenotypes. Therefore, we moved to ancestry prediction using the 1000 Genomes dataset which features fewer samples but higher population diversity. By splitting the data by sample superpopulations we achieved high inter-node heterogeneity. We showed that in this setting, frequent communication between the server and the clients plays a crucial role in achieving fast convergence and showing performance similar to that of the centralized model. We also demonstrated that depending on whether computational time or communication is a bottleneck of the system, FedAvg with different numbers of epochs in a round should be preferred.

In both of our experiments, the main reason for federated models not reaching the performance level of centralized models is data heterogeneity across the nodes, also called client dissimilarity. When a federated model trains on a client, it overfits to local data; then as fitted parameters from different clients get aggregated, the result may differ from the update of the corresponding centralized model, a phenomenon called client drift. Client drift can be decreased by increasing communication between the client and the server by decreasing the number of epochs of local training in a communication round (between parameter updates), as shown in Figures 4A, 5A. Another factor influencing client drift when training on heterogeneous data is the number of nodes, and correspondingly the degree of fragmentation of the data between the nodes, as federated training can become less stable with a larger number of nodes (Li et al., 2019). We did not observe this in our experiments, likely because of the genetic homogeneity of the White British UK Biobank participants and the low number of nodes in the 1000 Genomes experiment.

Federated learning is a quickly developing field of research and new strategies continue to emerge, including those aiming to tackle client drift and improve convergence in case of high inter-node heterogeneity, such as SCAFFOLD (Karimireddy et al., 2020) and FedDyn (Acar et al., 2021). These novel strategies make the training process more stable and may be preferable if communication between the clients and the server is limited or when the number of heterogeneous clients is large. However, they require additional testing, as FedAvg with frequent communication, one or two epochs in a round, is a difficult baseline to beat.

In both experiments, we used a single dataset artificially split into several independent nodes. On the one hand, this is an advantage as the uniform data collection process allows us to limit the influence of environmental and experimental confounders and focus on the relative performance of the models. On the other hand, in a real scenario, using multiple independently collected datasets may require additional work to unify features and outputs across datasets. For example, predictions from genomic data may require SNP imputation or another solution if different datasets have different sets of genetic variants; similarly, ancestry and phenotypes may be defined or collected differently in different experiments. Another issue that may be encountered in a real

scenario is that if independent datasets are isolated in trusted research environments, their joint analysis requires these TREs to allow sending information packets back and forth. In case of federated learning these packets would be relatively small (the size of the model) but frequent.

Federated learning enables data collaboration in genomics which may help solve several important problems. First, combining multiple datasets increases sample size which improves overall model accuracy and enables the prediction of rare diseases and inclusion of rare variants, which typically have larger effect sizes (Bodmer and Bonilla, 2008). Second, the vast majority of healthcare data currently comes from people of European descent, which makes models trained on this data biased toward Europeans, adding to the healthcare inequality of people around the world (Genetics, 2019; Martin et al., 2019b). Federated learning allows to include smaller datasets of different ancestries in the analysis and, thus, reduce the bias.

Being one of the first papers to explore federated learning on genomic data, this study has a limited scope. First, here we assume the trustworthiness of the parties. This is a reasonable assumption if data custodians, such as biobanks, give access to data upon application approval. However, data collaboration between different entities may require implementing additional privacy-enhancing mechanisms, as federated learning does not fully protect against privacy leakage. Second, we used basic machine learning models and the standard FedAvg strategy to keep things simple and focus on the relative performance of federated vs. centralized vs. local models. Depending on a specific problem, fine-tuned models may yield higher absolute performance; other FL strategies may achieve faster convergence with less communication. Third, as mentioned previously, establishing a real data collaboration may involve additional work to harmonize features and outputs between parties. Fourth, here we focused only on building a single "global" model, whereas some parties in data collaboration may require "personalized" models that prioritize their data (Kulkarni et al., 2020). We hope that these issues will be addressed in future studies.

5 Conclusions

Despite its promise, the applicability of federated learning to individual-level genomic data has not been sufficiently investigated. We filled this gap by training and analyzing federated models in two important scenarios: phenotype-from-genotype prediction and ancestry-from-genotype prediction. In the first experiment, on the UK Biobank data, we mimicked the scenario where data is collected within a single country and is distributed across data collection centers, for example, different hospitals. In the second experiment, on the 1000 Genomes data, we explored the potential of cross-continental collaboration and suggested how to maintain high performance despite high heterogeneity of the data. We showed that federated models consistently achieve high performance close to that of centralized models for the prediction of multiple phenotypes and ancestry, even in the presence of significant inter-node heterogeneity. For heterogeneous nodes, we investigated the dependency of federated models convergence on the amount of communication between the server and the nodes and provided recommendations on which schedule to choose if communication or computational time is a bottleneck. We also showed how federated prediction models can be integrated with federated data processing steps such as dimensionality reduction by federated PCA. This study encourages the adoption of federated models in healthcare, which has the potential to enable global data collaboration and train less biased models that represent diverse genetic ancestries.

Data availability statement

The data analyzed in this study is subject to the following licenses/restrictions: UK Biobank data is available upon request through the UK Biobank website. Requests to access this dataset should be directed to https://www.ukbiobank.ac.uk/enable-your-research/apply-for-access.

Ethics statement

Ethical approval was not required for the study involving humans in accordance with the local legislation and institutional requirements. Written informed consent to participate in this study was not required from the participants or the participants' legal guardians/next of kin in accordance with the national legislation and the institutional requirements.

Author contributions

DK: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Supervision, Visualization, Writing - original draft, Writing - review & editing. SM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing. AM: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Visualization, Writing - original draft, Writing - review & editing. ML: Investigation, Methodology, Visualization, Writing - review & editing. EK: Investigation, Software, Writing - review & editing. RV: Funding acquisition, Project administration, Resources, Writing - review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This study received funding from GENXT LTD. The funder was involved in the study design, collection, analysis, interpretation of data, the writing of this article and the decision to submit it for publication.

Acknowledgments

The computations performed in this study on the UK Biobank dataset were done locally on the Zhores cluster (Zacharov et al., 2019) and we thank the CDISE HPC team for their assistance. This research has been conducted using the UK Biobank Resource under

Application Number 43661. We also thank Dmitry Yarotsky, Pavel Nikonorov, and Bert Mouler for the valuable discussions.

Conflict of interest

All authors were employed by GENXT LTD.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2024.1266031/full#supplementary-material

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. (2016). "Deep learning with differential privacy," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (New York, NY), 308–318.

Acar, D. A. E., Zhao, Y., Navarro, R. M., Mattina, M., Whatmough, P. N., and Saligrama, V. (2021). Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*. doi: 10.48550/arXiv.2111.04263

Agrawal, A., Chiu, A. M., Le, M., Halperin, E., and Sankararaman, S. (2020). Scalable probabilistic PCA for large-scale genetic variation data. *PLoS Genet.* 16, e1008773. doi: 10.1371/journal.pgen.1008773

Alvarellos, M., Sheppard, H. E., Knarston, I., Davison, C., Raine, N., Seeger, T., et al. (2023). Democratizing clinical-genomic data: how federated platforms can promote benefits sharing in genomics. *Front. Genet.* 13, 1045450. doi: 10.3389/fgene.2022.1045450

Amariuta, T., Ishigaki, K., Sugishita, H., Ohta, T., Koido, M., Dey, K. K., et al. (2020). Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements. *Nat. Genet.* 52, 1346–1354. doi: 10.1038/s41588-020-00740-8

Beutel, D. J., Topal, T., Mathur, A., Qiu, X., Fernandez-Marques, J., Gao, Y., et al. (2020). Flower: a friendly federated learning research framework. *arXiv preprint arXiv:2007.14390*. doi: 10.48550/arXiv.2007.14390

Bodmer, W., and Bonilla, C. (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* 40, 695–701. doi: 10.1038/ng.f.136

Bonomi, L., Huang, Y., and Ohno-Machado, L. (2020). Privacy challenges and research opportunities for genomic data sharing. *Nat. Genet.* 52, 646–654. doi: 10.1038/s41588-020-0651-0

Buniello, A., MacArthur, J. A. L., Cerezo, M., Harris, L. W., Hayhurst, J., Malangone, C., et al. (2019). The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 47, D1005–D1012. doi: 10.1093/nar/gky1120

Chang, C. C., Chow, C. C., Tellier, L. C., Vattikuti, S., Purcell, S. M., and Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7. doi: 10.1186/s13742-015-0047-8

Chapman, C. R., Mehta, K. S., Parent, B., and Caplan, A. L. (2020). Genetic discrimination: emerging ethical challenges in the context of advancing technology. *J. Law Biosci.* 7, lsz016. doi: 10.1093/jlb/lsz016

Chen, Y., Qin, X., Wang, J., Yu, C., and Gao, W. (2020). Fedhealth: a federated transfer learning framework for wearable healthcare. *IEEE Intell. Syst.* 35, 83–93. doi: 10.1109/MIS.2020.2988604

Damgård, I., Pastro, V., Smart, N., and Zakarias, S. (2012). "Multiparty computation from somewhat homomorphic encryption," in *Annual Cryptology Conference* (Berlin; Heidelberg: Springer), 643–662.

Durvasula, A., and Lohmueller, K. E. (2021). Negative selection on complex traits limits phenotype prediction accuracy between populations. *Am. J. Hum. Genet.* 108, 620–631. doi: 10.1016/j.ajhg.2021.02.013

Evangelou, E., and Ioannidis, J. (2013). Meta-analysis methods for genome-wide association studies and beyond. *Nat. Rev. Genet.* 14, 379–389. doi: 10.1038/nrg3472

Falcon, W., Borovec, J., Walchli, A., Eggert, N., Schock, J., Jordan, J., et al. (2020). *PyTorchLightning/Pytorch-Lightning: 0.7.6 Release*.

Genetics (2019). Genetics for all. *Nat. Genet.* 51, 579. doi: 10.1038/s41588-019-0394-y

Gurdasani, D., Barroso, I., Zeggini, E., and Sandhu, M. S. (2019). Genomics of disease risk in globally diverse populations. *Nat. Rev. Genet.* 20, 520–535. doi: 10.1038/s41576-019-0144-0

Hartebrodt, A., Nasirigerdeh, R., Blumenthal, D. B., and Röttger, R. (2021). "Federated principal component analysis for genome-wide association studies," in *2021 IEEE International Conference on Data Mining (ICDM)* (New York, NY: IEEE), 1090–1095.

Hartebrodt, A., and Röttger, R. (2022). Federated horizontally partitioned principal component analysis for biomedical applications. *Bioinform. Adv.* 2, vbac026. doi: 10.1093/bioadv/vbac026

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning With Sparsity: The Lasso and Generalizations*. Boca Raton, FL: CRC Press.

Joly, Y., Feze, I. N., Song, L., and Knoppers, B. M. (2017). Comparative approaches to genetic discrimination: chasing shadows? *Trends Genet.* 33, 299–302. doi: 10.1016/j.tig.2017.02.002

Joshi, M., Pal, A., and Sankarasubbu, M. (2022). Federated learning for healthcare domain-pipeline, applications and challenges. *ACM Trans. Comput. Healthcare* 3, 40. doi: 10.1145/3533708

Jung, A. (2022). *Machine Learning*. Singapore: Springer Nature.

Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S., Stich, S., and Suresh, A. T. (2020). "Scaffold: stochastic controlled averaging for federated learning," in *International Conference on Machine Learning* (PMLR), 5132–5143.

Kavianpour, S., Sutherland, J., Mansouri-Benssassi, E., Coull, N., and Jefferson, E. (2022). Next-generation capabilities in trusted research environments: interview study. *J. Med. Internet Res.* 24, e33720. doi: 10.2196/33720

Kirkpatrick, B. E., and Rashkin, M. D. (2017). Ancestry testing and the practice of genetic counseling. *J. Genet. Counsel.* 26, 6–20. doi: 10.1007/s10897-016-0014-2

Kulkarni, V., Kulkarni, M., and Pant, A. (2020). "Survey of personalization techniques for federated learning," in *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)* (IEEE), 794–797.

Lambert, S. A., Gil, L., Jupp, S., Ritchie, S. C., Xu, Y., Buniello, A., et al. (2021). The polygenic score catalog as an open database for reproducibility and systematic evaluation. *Nat. Genet.* 53, 420–425. doi: 10.1038/s41588-021-00783-5

Lello, L., Raben, T. G., Yong, S. Y., Tellier, L. C., and Hsu, S. D. (2019). Genomic prediction of 16 complex disease risks including heart attack, diabetes, breast and prostate cancer. *Sci. Rep.* 9, 1–16. doi: 10.1038/s41598-019-51258-x

Lewis, C. M., and Vassos, E. (2020). Polygenic risk scores: from research tools to clinical instruments. *Genome Med.* 12, 1–11. doi: 10.1186/s13073-020-00742-5

Li, X., Gu, Y., Dvornek, N., Staib, L. H., Ventola, P., and Duncan, J. S. (2020). Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: abide results. *Med. Image Anal.* 65, 101765. doi: 10.1016/j.media.2020.101765

Li, X., Huang, K., Yang, W., Wang, S., and Zhang, Z. (2019). On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*. doi: 10.48550/arXiv.1907.02189

Lim, J. Q., and Chan, C. S. (2021). "From gradient leakage to adversarial attacks in federated learning," in *2021 IEEE International Conference on Image Processing (ICIP)*, 3602–3606.

Manichaikul, A., Mychaleckyj, J. C., Rich, S. S., Daly, K., Sale, M., and Chen, W.-M. (2010). Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867–2873. doi: 10.1093/bioinformatics/btq559

Mansouri-Benssassi, E., Rogers, S., Smith, J., Ritchie, F., Jefferson, E., Scotland, N., et al. (2021). Machine learning models disclosure from trusted research environments (TRE), challenges and opportunities. *arXiv preprint arXiv:2111.05628*. doi: 10.48550/arXiv.2111.05628

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019a). Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* 51, 584–591. doi: 10.1038/s41588-019-0379-x

Martin, A. R., Kanai, M., Kamatani, Y., Okada, Y., Neale, B. M., and Daly, M. J. (2019b). Current clinical use of polygenic scores will risk exacerbating health disparities. *Nat. Genet.* 51, 584.

Mayo, K. R., Basford, M. A., Carroll, R. J., Dillon, M., Fullen, H., Leung, J., et al. (2023). The all of us data and research center: Creating a secure, scalable, and sustainable ecosystem for biomedical research. *Annu. Rev. Biomed. Data Sci.* 6, 443–464. doi: 10.1146/annurev-biodatasci-122120-104825

McMahan, B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. (2017). "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics* (PMLR), 1273–1282.

Mo, F., Haddadi, H., Katevas, K., Marin, E., Perino, D., and Kourtellis, N. (2021). "PPFL: privacy-preserving federated learning with trusted execution environments," in *MobiSys '21: Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services* (ACM), 94–108.

Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., and Srivastava, G. (2021). A survey on security and privacy of federated learning. *Future Gen. Comput. Syst.* 115, 619–640. doi: 10.1016/j.future.2020.10.007

Nasirigerdeh, R., Torkzadehmahani, R., Matschinske, J., Frisch, T., List, M., Späth, J., et al. (2020). sPLINK: a federated, privacy-preserving tool as a robust alternative to meta-analysis in genome-wide association studies. *bioRxiv*. doi: 10.1101/2020.06.05.136382

Nik-Zainal, P. S., Seeger, T., Fennessy, R., Hall, E., Moss, P., Coles, G., et al. (2022). *Multi-party Trusted Research Environment Federation: Establishing Infrastructure for Secure Analysis Across Different Clinical-Genomic Datasets*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). "PyTorch: an imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32* (Curran Associates, Inc.), 8024–8035.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genet.* 2, e190. doi: 10.1371/journal.pgen.0020190

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38, 904–909. doi: 10.1038/ng1847

Privé, F., Aschard, H., Carmi, S., Folkersen, L., Hoggart, C., O'Reilly, P. F., et al. (2022). Portability of 245 polygenic scores when derived from the UK Biobank and applied to 9 ancestry groups from the same cohort. *Am. J. Hum. Genet.* 109, 12–23. doi: 10.1016/j.ajhg.2021.11.008

Prive, F., Aschard, H., Ziyatdinov, A., and Blum, M. G. (2018). Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinformatics* 34, 2781–2787. doi: 10.1093/bioinformatics/bty185

Privé, F., Luu, K., Blum, M. G., McGrath, J. J., and Vilhjálmsson, B. J. (2020). Efficient toolkit implementing best practices for principal component analysis of population genetic data. *Bioinformatics* 36, 4449–4457. doi: 10.1093/bioinformatics/btaa520

Qian, J., Tanigawa, Y., Du, W., Aguirre, M., Chang, C., Tibshirani, R., et al. (2020). A fast and scalable framework for large-scale and ultrahigh-dimensional sparse regression with application to the UK Biobank. *PLoS Genet.* 16, e1009141. doi: 10.1371/journal.pgen.1009141

Ray, D., and Boehnke, M. (2018). Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genet. Epidemiol.* 42, 134–145. doi: 10.1002/gepi.22105

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., et al. (2020). The future of digital health with federated learning. *NPJ Digit. Med.* 3, 1–7. doi: 10.1038/s41746-020-00323-1

Sadat, M. N., Al Aziz, M. M., Mohammed, N., Chen, F., Jiang, X., and Wang, S. (2018). Safety: secure GWAS in federated environment through a hybrid solution. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 16, 93–102. doi: 10.1109/TCBB.2018.2829760

Shi, H., Gazal, S., Kanai, M., Koch, E. M., Schoech, A. P., Siewert, K. M., et al. (2021). Population-specific causal disease effect sizes in functionally important regions impacted by selection. *Nat. Commun.* 12, 1–15. doi: 10.1038/s41467-021-21286-1

Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., et al. (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat. Genet.* 53, 185–194. doi: 10.1038/s41588-020-00757-z

The 1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature* 526, 68. doi: 10.1038/nature15393

The Global Alliance for Genomics and Health (2016). A federated ecosystem for sharing genomic, clinical data. *Science* 352, 1278–1280. doi: 10.1126/science.aaf6162

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58, 267–288.

UK Health Data Research Alliance and NHSX (2021). *Building Trusted Research Environments - Principles and Best Practices; Towards TRE ecosystems.*

Vaid, A., Jaladanki, S. K., Xu, J., Teng, S., Kumar, A., Lee, S., et al. (2021). Federated learning of electronic health records to improve mortality prediction in hospitalized patients with covid-19: machine learning approach. *JMIR Med. Inform.* 9, e24207. doi: 10.2196/24207

Wjst, M. (2010). Caught you: threats to confidentiality due to the public release of large-scale genetic data sets. *BMC Med. Ethics* 11, 21. doi: 10.1186/1472-6939-11-21

Xu, J., Glicksberg, B. S., Su, C., Walker, P., Bian, J., and Wang, F. (2021). Federated learning for healthcare informatics. *J. Healthcare Inform. Res.* 5, 1–19. doi: 10.1007/s41666-020-00082-4

Yang, H.-C., Chen, C.-W., Lin, Y.-T., and Chu, S.-K. (2021). Genetic ancestry plays a central role in population pharmacogenomics. *Commun. Biol.* 4, 1–14. doi: 10.1038/s42003-021-01681-6

Yang, Q., Liu, Y., Chen, T., and Tong, Y. (2019). Federated machine learning: concept and applications. *ACM Trans. Intell. Syst. Technol.* 10, 1–19. doi: 10.1145/3298981

Zacharov, I., Arslanov, R., Gunin, M., Stefonishin, D., Bykov, A., Pavlov, S., et al. (2019). 'Zhores'—Petaflops supercomputer for data-driven modeling, machine learning and artificial intelligence installed in Skolkovo Institute of Science and Technology. *Open Eng.* 9, 512–520. doi: 10.1515/eng-2019-0059