FeatureCloud

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare

Deliverable D8.5
"Feedback on public data performance"

Work Package WP8
"Testing and evaluation in clinical translation"

Disclaimer

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078. Any dissemination of results reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Copyright message

Document information

Grant Agreement Number: 826078			**Acronym: FeatureCloud**	
Full title	Privacy preserving federated machine learning and blockchaining for reduced cyber risks in a world of distributed healthcare			
Topic	Toolkit for assessing and reducing cyber risks in hospitals and care centres to protect privacy/data/infrastructures			
Funding scheme	RIA - Research and Innovation action			
Start Date	1 January 2019	**Duration**	60 months	
Project URL	https://featurecloud.eu/			
EU Project Officer	Christos Maramis, Health and Digital Executive Agency (HaDEA)			
Project Coordinator	Jan Baumbach, University of Hamburg (UHAM)			
Deliverable	D8.5 - Feedback on public data performance			
Work Package	WP8 - Testing and evaluation in clinical translation			
Date of Delivery	**Contractual**	30/06/2023 (M54)	**Actual**	29/06/2023 (M54)
Nature	Report	**Dissemination Level**	Public	
Lead Beneficiary	01 UHAM			
Responsible Author(s)	Mohammad Bakhtiari (UHAM)			
	Walter Hötzendorfer (RI)			
Keywords	Federated Learning, Evaluation, Federated workflow			

History of changes

Version	Date	Contributions	Contributors (name and institution)
V0.1	01/06/2023	First draft	Mohammad Bakhtiari (UHAM)
V0.2	12/06/2023	Review by WP8 lead	Walter Hötzendorfer (RI)
V0.3	23/06/2023	Internal Review	Sándor Fejér (GND)
V1.0	28/06/2023	Final version	Mohammad Bakhtiari (UHAM)

Actual effort in person-months (PMs)

Contributor (name and institution)	Invested resources (deliverable)	Overview of contributions
Mohammad Bakhtiari (UHAM)	1.60 PM	First draft, final version
Walter Hötzendorfer (RI)	0.10 PM	Review by WP8 leader
Sándor Fejér (GND)	0.05 PM	Internal review

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

Table of Content

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

1 Table of acronyms and definitions

COPD	Chronic obstructive pulmonary disease
dbGaP	Database of Genotypes and Phenotypes
DP	Differential Privacy
EGA	European Genome-phenome Archive
eQTL	Expression quantitative trait loci
GEO	Gene Expression Omnibus (online platform)
GND	Gnome Design SRL
GTEx	Genotype-Tissue Expression Program
MUG	Medizinische Universität Graz
Patients	In this deliverable, we use the term “patients” for all research subjects. In FeatureCloud, we will focus on patients, as this is already the most vulnerable case scenario and this is where most primary data is available to us. Admittedly, some research subjects participate in clinical trials but not as patients but as healthy individuals, usually on a voluntary basis and are therefore not dependent on the physicians who care for them. Thus, to increase readability, we simply refer to them as “patients”.
RI	Research Institute AG & Co. KG
SBA	SBA Research Gemeinnützige GmbH
SDU	Syddansk Universitet
sFL	FL and additive secret sharing
SMPC	Secure Multi-Party Computation
SNP	Single Nucleotide Polymorphisms
TCGA	Cancer Genome Atlas Program
UHAM	Universität Hamburg
UMR	Philipps Universität Marburg
WP	Work package

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

2 Objectives of the deliverable based on the Description of Action (DoA)

Deliverable D8.5 “Feedback on public data performance” is related to task 4, “Evaluation” of WP7 as described in the Description of Action. In that regard, we implemented various automatized measures for different applications in the FeatureCloud platform to evaluate the performance of federated methods. D8.5 also touches on task 4 of WP8 by comparing performance of different FeatureCloud applications in federated scenarios on publicly available medical or biological datasets while addressing real world challenges like data heterogeneity, imbalanced-ness, batch-effect, etc.

WP7 - Integrated FeatureCloud health informatics platform and app store

Objective 4:

To implement automatized measures for evaluating the overall strategy of FeatureCloud by demonstrating that the performance of federated machine learning (in terms of accuracy) is comparable to the performance of traditional cloud-based approaches (Task 4)

Task 4: Evaluation

The federated machine learning paradigm will be validated and tested by using publicly available data, e.g. from The Cancer Genome Atlas and the Amsterdam Classification Evaluation Suite. We will first distribute this data over artificial hospitals with servers behind own firewalls. Afterwards, we will evaluate the prediction performance of the federated machine learning approaches against classical, non-federated tools (having centralized access to the full data in a traditional cloud solution). As a next step, original clinical trial data from WP8 (which was or will be anonymized and prepared for traditional cloud computing to comply with legal privacy requirements) will be used for second level testing.

WP8 - Testing and evaluation in clinical translation

Objective 3:

To evaluate the technical performance of the FeatureCloud platform in terms of accuracy and translational power in clinical settings by re-analysing data previously processed in a traditional cloud-based approach and to provide feedback to WPs 4, 5 and 7 (Task 4)

Task 4: Performance evaluation and translation into real-world clinical studies

This task is tightly connected to task 4 of WP7. All partners will beta test the FeatureCloud platform, while coordinating partner UHAM and partner GND (WP7) will account for the feedback by suggesting and implementing changes to the software design. Original clinical data will be used for evaluating and improving the applicability of the FeatureCloud platform in real world settings of clinical study practice. Such data is available to UHAM already through previous and ongoing projects. Patients’ consent and ethical approval already exist. The consortium will use the FeatureCloud platform to analyse data from a set of at least three clinical datasets. Based on this data, all partners will evaluate data processing capabilities and provide feedback to UHAM, GND (WP7), and UMR (WP3) regarding usability and feature requests, and to consortium partners MUG and SDU (WPs 3 and 4) regarding the performance of the federated vs. the standard methodology (also see task 1). In addition, RI will provide feedback regarding privacy (see task 6).

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

3 Executive Summary

This deliverable covers multiple peer-reviewed publications utilizing FeratureCloud as a federated platform to replicate the best achievable results in the centralized training using federated learning. In this way we acquire feedback from the research community on public data to show the accuracy and performance of our federated algorithms. Performance has been evaluated by us using different cross-validation schemes, and subsequently by many independent reviewers of several publications in peer reviewed journals. In that regard, different methodologies were implemented in federated fashion and applied on various public datasets (See section 4). For instance, Flimma (see section 4.1), as a federated version of Limma-voom was applied on Veteran (US Veterans' Administration lung cancer study data), Lung (NCCTG lung cancer data), and Rossi (Criminal recidivism) datasets (See section 5.1). Besides, Partea conducts time to event analysis on COPDGene chronic obstructive pulmonary disease datasets (See section 4.2 and 5.2), and sPLINK is applied on the "TCGA-BRCA" [6] dataset from the Cancer Genome Atlas Program (see section 4.3 and 5.3).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

4 Introduction (Challenge)

FeatureCloud [1] as a platform for federated learning keeps raw data on the local device, while minimizing data collection and processing on a centralized server. This approach provides a variety of privacy advantages out of the box. The privacy concerns associated with the use of federated learning serve to motivate the desire to keep raw data on each local device in a distributed machine learning setting. There are challenges in applying federated learning to biomedicine or biology studies ranging from availability of sufficient high-quality data to the requirements of validating a locally trained model on data from external sources.

Based on multiple peer-reviewed publications, we evaluated the performance FeratureCloud applications using different cross-validation schemes. We acquired feedback by many independent reviewers of several publications in peer reviewed journals which demonstrates that applying federated models in the FeatureCloud platform gives comparable results to centralized training. In that regard, we provide an overview of some of FeatureCloud applications that are published in peer-reviewed journals to show how they address different challenges ranging from privacy enhancing technologies to data heterogeneity.

Accordingly, applying various federated methods, while simultaneously utilizing secure multi-party computation (SMPC) or DP, on publicly available data from different fields, provides comparable results to the corresponding state of the art centralized approach according to the feedback from reviewers of multiple peer-reviewed publications.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

5 Methodology

In this section we elaborate on the evaluation strategy of applying FeatureCloud on publicly available data. In multiple peer-reviewed studies we have shown that prototypes trained by adopting federated learning strategies are able to achieve reliable performance [2, 3, 4], according to the feedback from the reviewers of the journals, thus generating robust models without sharing data and limiting the impact on security and privacy. The FeatureCloud consortium conducted research on public data from different biomedical domains and published the results in peer-reviewed journals. The public nature of the data used facilitates the reproducibility of results for the sake of comparison with alternative methodologies. In the following subsections, we describe the methodologies of these studies. Corresponding apps are available in the FeatureCloud App store.

5.1 Flimma

The process of identification of differentially expressed genes or transcripts is a critical task in molecular systems medicine, which involves comparing the gene expression profiles of two or more groups of samples to reveal genes with significant differences between the groups. High-throughput gene expression profiling technologies such as microarrays and RNA sequencing are used to identify differentially expressed genes. However, these technologies have their biases, and the results obtained from each platform may be different. Several bioinformatics tools have been developed to identify differentially expressed genes from such data. These methods differ in the assumptions about data distribution, data normalization strategies, and the test statistic used to detect differentially expressed genes. One significant challenge of differential expression studies is the lack of robustness due to the high technical and biological variability of the data. Many strategies can be used to address this, including increasing the sample size, which is non-trivial as data collection is expensive and time-consuming, sample availability may be limited, and existing data may not be shareable due to personal data protection laws.

Privacy issues are also a significant concern in differential expression studies. The statistical analysis of expression data may require relevant clinical metadata, which may be identified when combined. Recent works suggest that patient genotypes can be predicted from RNA-seq data, making patients identifiable through expression profiles or so-called “expression quantitative trait loci” (eQTL) data obtained from open-access sources. To control the exchange of sensitive molecular profiling data, databases, such as dbGaP or EGA [21], restrict access to authorized users affiliated with organizations willing to guarantee the legal and secure use of personal data. Alternatively, researchers can combine the results of several studies using meta-analysis techniques such as Fisher’s method, Stouffer’s method, RankProd, or the random effects model. However, the main disadvantage of meta-analysis tools is that their underlying assumptions about the distribution of p-values or effect sizes may not be realistic, and they may ignore possible differences between cohorts or data processing steps that may significantly impact the results.

Privacy-aware techniques such as federated learning (FL), differential privacy (DP), homomorphic encryption (HE), and secure multi-party computation (SMPC) have recently moved into the focus of research for tasks involving privacy-sensitive patient data. FL implies collaborative model training by multiple participants without disclosing private data to any other party. DP perturbs the data or results by adding noise to them. HE performs computation on the encrypted data from the participants. SMPC computes secret shares from the data and shares them with the computing parties. FL is a promising alternative to SMPC and HE in terms of performance and scalability. The privacy of federated methods can be enhanced by applying HE or SMPC on the shared model parameters.

Flimma [2] (federated limma) is a novel federated privacy-aware tool for the identification of differentially expressed genes. It represents a federated implementation of the popular differential expression analysis workflow limma voom. Flimma is based on HyFed [16], a hybrid FL framework,

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 826078.

which applies additive secret sharing-based SMPC method to avoid disclosing the local model parameters to the server. It provides several advantages over the existing approaches for gene expression analysis, including enhancing the privacy of the data in the cohorts since the expression profiles never leave the local execution sites, and only aggregated parameters are revealed to the server and the other local sites. Flimma is particularly robust against heterogeneous distributions of data, making it a powerful alternative for multi-center studies where patient privacy is a key concern.

5.1.1 Datasets

The Flimma tool implements a federated version of the limma voom workflow and is a privacy-aware tool for differential expression analysis. Flimma is designed to operate on distributed cohorts without disclosing sensitive data, and it uses a hybrid federated approach where the local parameters of the clients are hidden from the server, and only global parameters resulting from the aggregation are disclosed. Flimma has been tested on two real-world datasets: a breast cancer expression dataset from the Cancer Genome Atlas Program (TCGA [6]) and a skin dataset from the Genotype-Tissue Expression Program (GTEx [17]). Flimma is robust to technical batch effects, and it models the batch effects of datasets by adding m−1 binary covariates to the linear model, where m is the number of datasets. Meanwhile, Flimma is applied to three additional publicly available breast cancer cohorts from the Gene Expression Omnibus (GEO): GSE129508 [7], GSE149276 [8], and GSE58135 [9].

Analyzing a federated model on public data is important because it enables researchers to test and validate the performance of the model on data that is not from the same source as the original data. This can help to identify any biases or limitations in the model, and it can also help to improve the generalizability of the model. In the case of Flimma, it is not clear from the context which public datasets it was analyzed on. However, the authors of the paper note that one limitation of their work is the absence of a gold standard for the evaluation of differential expression analysis results. Therefore, analyzing Flimma on public datasets could be a useful way to evaluate its performance and compare it to other methods for normalization and differential expression analysis.

The Genotype-Tissue Expression (GTEx) program is a comprehensive data resource and tissue bank that aims to study the relationship between genetic variation and gene expression across multiple human tissues and individuals. The GTEx program has created a reference dataset to study genetic changes and gene expression, and has generated a large dataset that includes over 10,000 bulk RNA-seq samples. The GTEx program has also established a comprehensive catalog of genetic variants that affect gene expression across multiple tissues, called expression quantitative trait loci (eQTLs).

The skin dataset from GTEx is part of the larger GTEx dataset, which includes RNA-seq data from 54 different tissue sites. The skin dataset includes RNA-seq data from skin tissue samples collected from donors in the GTEx program. The skin dataset can be used to study gene expression patterns in skin tissue, as well as the relationship between genetic variants and gene expression in skin tissue.

The GTEx program has generated a large dataset that includes over 10,000 bulk RNA-seq samples from multiple human tissues and individuals. The RNA-seq data in the GTEx dataset is generated using Illumina HiSeq platform and is available in gene-level expression format. The GTEx dataset also includes genotype data from approximately 948 post-mortem donors and RNA-seq data from approximately 17,382 samples across 54 tissue sites and 2 cell lines. Full gene expression datasets are available for download through the GTEx Portal, while genotypes and RNA-seq bam files are available via the database of Genotypes and Phenotypes (dbGaP).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

5.1.2 Method

Flimma is designed to operate on distributed cohorts without the disclosure of sensitive data and employs a hybrid federated approach to enhance data privacy. The workflow of Flimma includes several steps, including filtering genes with insufficient counts, performing UQ normalization, fitting linear regression models, and computing p-values, fold-changes, and moderated t statistics for each gene. Flimma is implemented as a federated version of the limma voom workflow, and each Flimma client accepts a matrix of read counts and a design matrix specifying class labels and covariates for each sample. Flimma is publicly available and is a promising alternative to meta-analysis methods for multi-center gene expression projects.

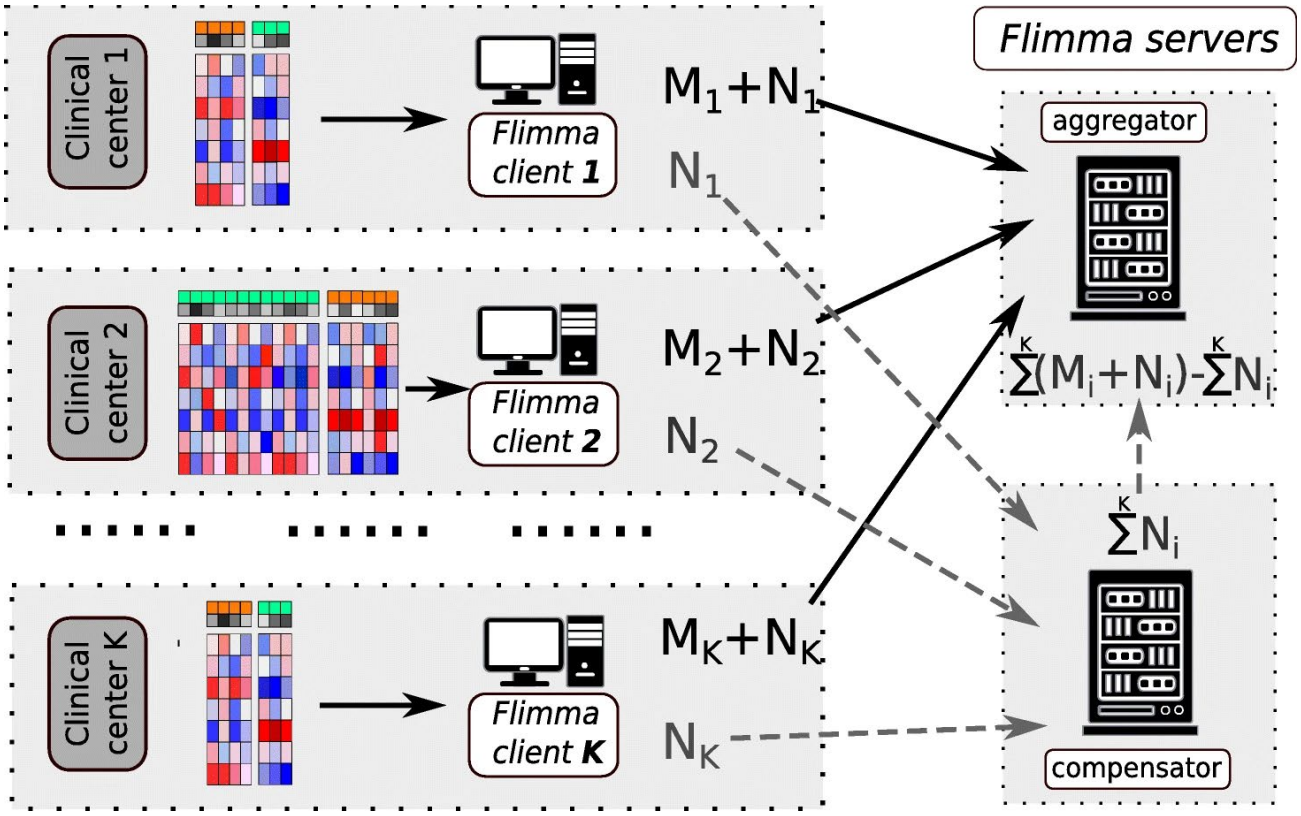

Figure 1. The scheme of Flimma: M denotes local intermediate parameters, N denotes local noise. K is the total number of participants [2].

Flimma is a privacy-aware tool for differential expression analysis that uses a hybrid federated approach to enhance data privacy. Flimma is based on HyFed [5], a hybrid federated learning (FL) framework that applies additive secret sharing-based secure multi-party computation (SMPC) to avoid disclosing the local model parameters to the server.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

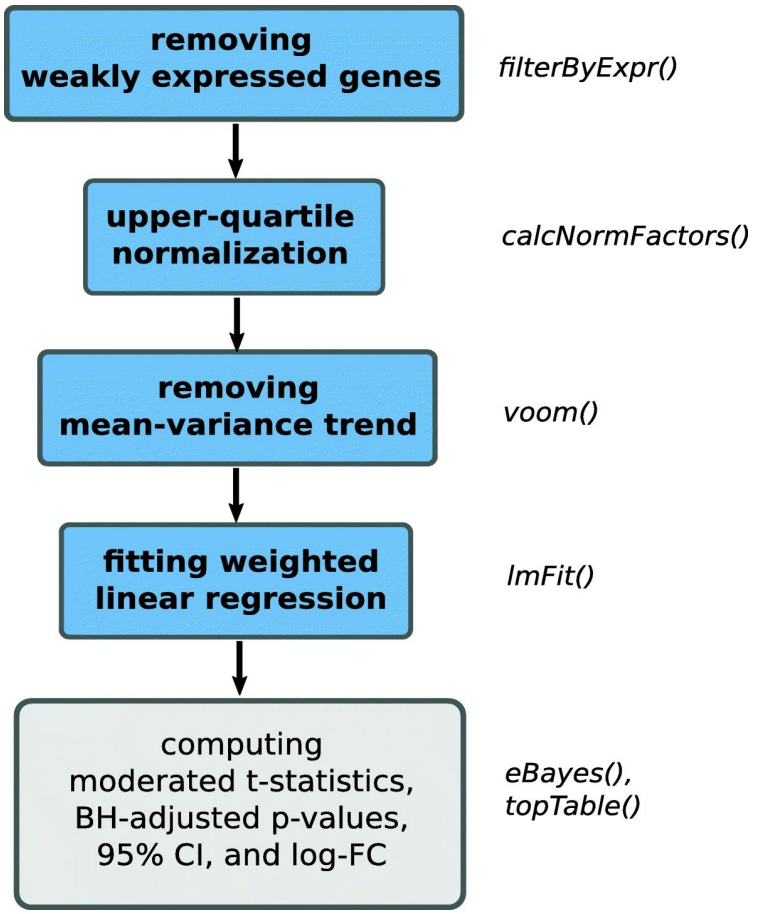

Figure 2. The scheme of the Flimma workflow. Steps that were reimplemented in a federated fashion are shown in blue. The names of the functions used in the limma voom workflow are shown on the right of the flowchart [2].

5.2 PARTEA

PARTEA (Privacy-Aware Real-Time Event Analysis) is an advanced framework that combines the principles of privacy protection and real-time event analysis. PARTEA is designed to address the challenges of analyzing time-dependent data while ensuring the privacy of sensitive information. In privacy-aware time-to-event analysis, the focus is on studying the time it takes for certain events to occur while respecting the privacy of individuals or organizations involved. This type of analysis is often relevant in various domains such as healthcare, finance, and social sciences, where understanding the time-to-event relationships is crucial for decision-making and prediction.

PARTEA employs sophisticated algorithms and techniques to analyze time-to-event data in real-time, extracting valuable insights without compromising privacy. It utilizes privacy-preserving methodologies like differential privacy, secure multiparty computation, or anonymization techniques to protect the identities and sensitive attributes of individuals or entities involved in the analysis.

By combining privacy protection with real-time event analysis, PARTEA enables researchers, analysts, and organizations to derive meaningful and timely insights from time-dependent data while upholding privacy regulations and ethical considerations. This framework opens up new avenues for research, decision-making, and innovation, ensuring the balance between data-driven analysis and privacy preservation in a rapidly evolving digital landscape.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

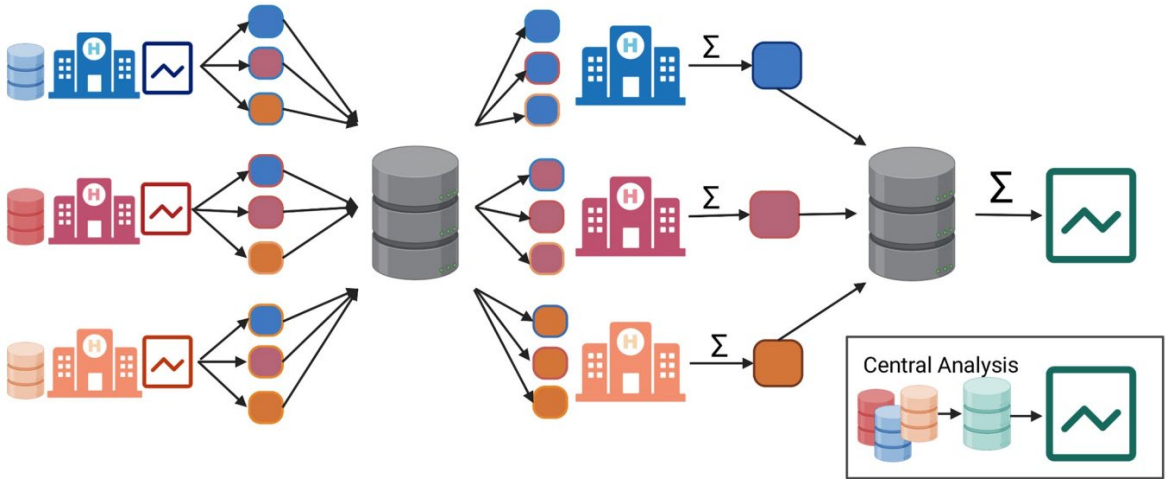

Figure 3. Hybrid federated learning workflow using additive secret sharing. Each institution calculates its local statistics and creates a secret for each participant (1). The global aggregation server receives the secrets and distributes them to the corresponding participants (2). Each local client decrypts the secrets and sums them up (3). The sum is shared with the global aggregation server (4), which sums them up again, revealing the final global aggregation (5). Created with Biorender.com [3].

5.2.1 Dataset

Partea uses three different datasets to experiment on:

- Veteran (US Veterans' Administration lung cancer study data) [18]: This dataset contains information on 137 patients with lung cancer who were treated at the Veterans' Administration Medical Center in West Los Angeles between 1969 and 1971. The dataset can be accessed through the following link: https://biostat.app.vumc.org/wiki/Main/DataSets

- Lung (NCCTG lung cancer data) [18]: This dataset contains information on 168 patients with advanced non-small cell lung cancer who participated in a clinical trial conducted by the North Central Cancer Treatment Group (NCCTG) between 1980 and 1983. The dataset can be accessed through the following link: https://biostat.app.vumc.org/wiki/Main/DataSets

- Rossi (Criminal recidivism data) [19]: This dataset contains information on 432 male offenders who were released from prison in Michigan between 1965 and 1974. The dataset includes variables such as age, race, marital status, prior criminal record, and whether or not the offender was employed at the time of release. The dataset can be accessed through the following link: https://www.rand.org/pubs/reports/R1057.html

5.3 sPLINK

sPLINK, in the context of genome-wide association studies (GWAS), refers to a hybrid federated tool that serves as a robust alternative to traditional meta-analysis. GWAS involves studying the genetic variations across a large number of individuals to identify associations between specific genetic variants and diseases or traits.

Meta-analysis is a commonly used approach in GWAS, where data from multiple studies are combined to increase statistical power and detect genetic associations that may not be significant in individual studies. However, meta-analysis requires sharing individual-level genetic data, which can raise privacy concerns and encounter legal or ethical barriers.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

sPLINK (secure PLINK) offers a privacy-preserving solution for conducting GWAS by leveraging federated learning techniques. Federated learning allows collaboration and analysis across multiple institutions or datasets without sharing raw data. In the context of sPLINK, each participating institution retains control of its data while contributing aggregated statistics or model updates to the overall analysis.

By using sPLINK, researchers can perform joint analysis on GWAS datasets without directly accessing or sharing sensitive genetic information. This hybrid federated approach ensures privacy protection and addresses the challenges associated with data sharing in large-scale genetic studies. Additionally, sPLINK maintains the statistical power of traditional meta-analysis, making it a promising alternative for conducting robust and privacy-aware GWAS.

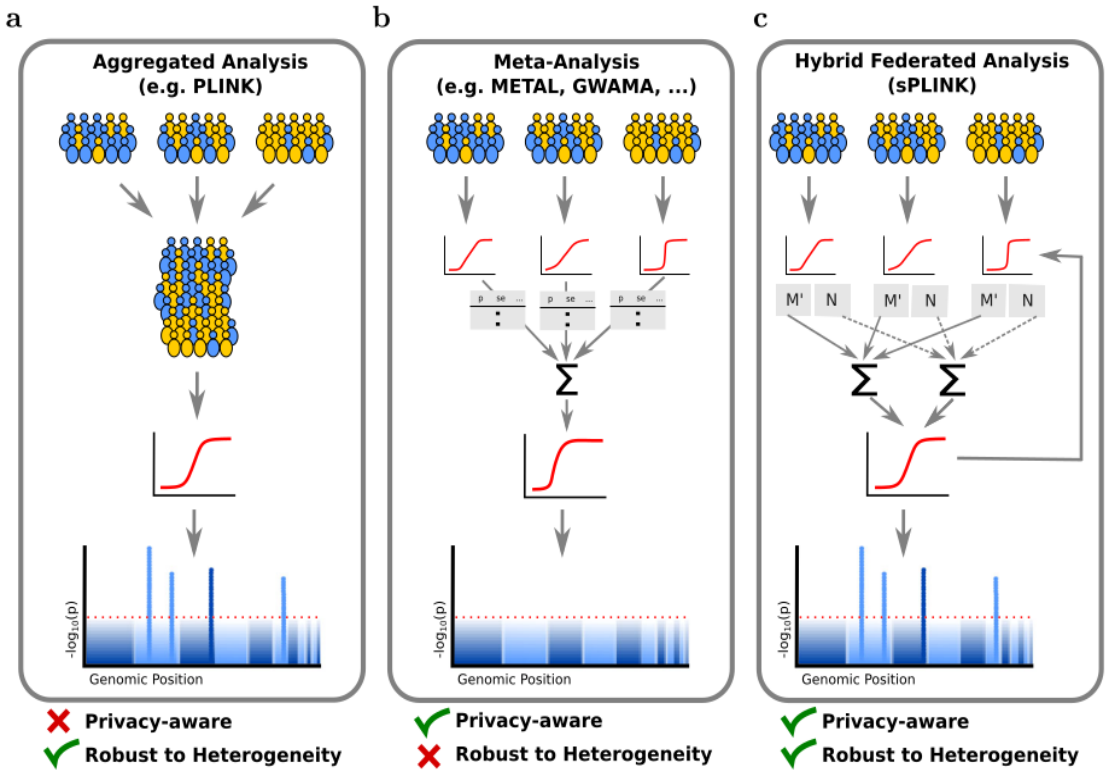

Figure 4. Comparison of sPLINK (c), aggregated analysis (a), and meta-analysis (b) approaches: Aggregated analysis requires cohorts to pool their private data for a joint analysis. The meta-analysis approaches aggregate the summary statistics from the cohorts to estimate the combined p-values. In sPLINK, the cohorts calculate the model parameters (M) from the local data and global model, generate noise (N), and make the parameters noisy (M) in an iterative manner. The aggregated noise and noisy parameters are in turn aggregated to update the global model or build the final model. sPLINK combines the advantages of the aggregated analysis and meta-analysis, i.e., robustness against heterogeneous data and enhancing the privacy of cohorts' data. Yellow/blue color indicates case/control samples.

5.3.1 Dataset

sPLINK experiments with the COPDGene dataset (http://www.copdgene.org/) [20],. This is a publicly available dataset that contains genetic and clinical data from individuals with chronic obstructive pulmonary disease (COPD) and controls without COPD. It is available through the dbGaP accession number "phs000179.v1.p1".

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

6 Results

After conducting federated analysis on public data using the FeatureCloud platform, we analyzed the results in terms of comparing the performance of the federated application with the centralized model considering various real world scenarios and challenges. The detailed performance analysis which considers the feedback from the reviewers of the journals, is explained in the following subsections for three different peer-reviewed publications [2, 3, 4].

6.1 Flimma

Flimma experimented with the GTEx dataset by applying its federated privacy-aware tool for differential expression analysis on the dataset. The GTEx dataset used in Flimma includes 1277 skin expression profiles with sun exposure as the target class label and patient age and sex as covariates. Flimma tested its power by modeling the multi-party setting through randomly partitioning the dataset into virtual cohorts while introducing different levels of imbalance with respect to target class labels and covariate distributions. Flimma simulated three realistic scenarios leading to different levels of sample distribution heterogeneity between local cohorts to assess its power.

In Flimma's experiments, the GTEx dataset was split into virtual cohorts to simulate a federated scenario. The dataset was partitioned randomly while introducing different levels of imbalance with respect to target class labels and covariate distributions. Flimma simulated three realistic scenarios leading to different levels of sample distribution heterogeneity between local cohorts. This allowed Flimma to test its power in a multi-party setting and assess its ability to operate on distributed cohorts without disclosing sensitive data.

6.1.1 Imbalanced Scenario

Flimma and meta-analysis approaches are compared on skin data from GTEx. Flimma's power is assessed by partitioning the datasets into virtual cohorts with varying class label imbalance and covariate distributions. They simulated realistic scenarios to represent heterogeneity in sample distribution between local cohorts. The limma voom results on the pooled datasets is considered as gold standard. In summary, Flimma obtained the same results as limma voom in all tests. Across all experiments, the maximal absolute difference for log-transformed p-values and log-fold-change values computed by Flimma and limma voom did not exceed 0.1. In contrast, the results of the meta-analysis methods diverged from the results of limma voom, and this effect was especially pronounced in imbalanced scenarios.

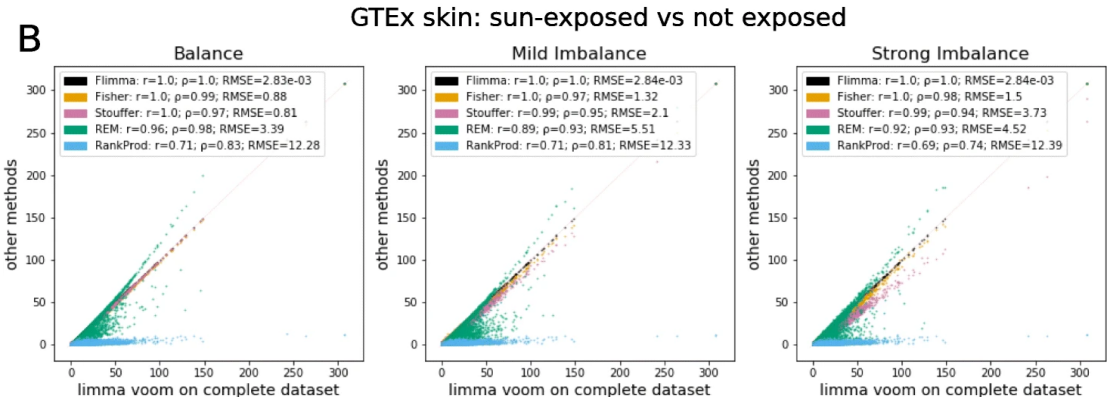

Figure 5. The comparison of negative log-transformed p-values computed by Flimma and meta-analysis methods (y-axis) with p-values obtained by limma on the aggregated dataset (x-axis) in three scenarios on GTEx skin datasets. Pearson correlation coefficient (r), Spearman correlation coefficient (ρ), and root-mean squared error (RMSE) calculated for each method are reported in the legend [2].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

6.1.2 Performance on Top-ranked genes

One of the key indicators in performance comparison of Flimma with Meta-analysis approaches is identification of a small number of significantly differentially expressed genes. By investigating the performance of the methods regarding the various numbers of selected top differentially expressed genes after sorting by *p*-value, it is shown that *Flimma* perfectly reproduced the results of aggregated *limma voom* in all scenarios and outperformed all meta-analysis approaches. Fisher's and Stouffer's methods demonstrated almost perfect performance in the balanced scenario, but their performance decreased in the imbalanced ones.

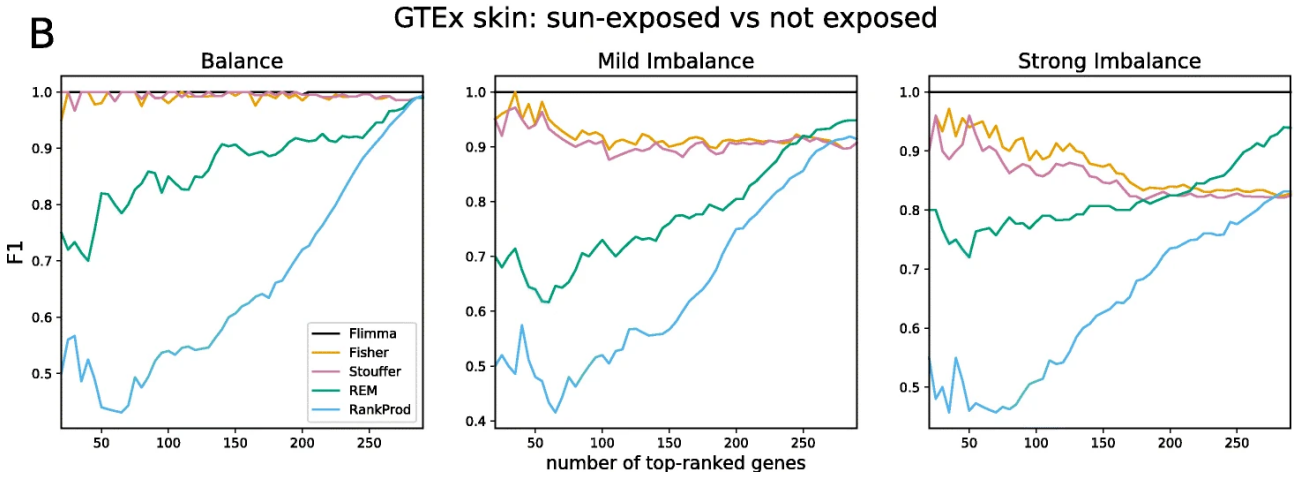

Figure 6. The dependency of the F1 score on the number of top-ranked genes considered to be differentially expressed. Genes were ranked in order of their negative log-transformed p-values decreasing and the number of top-ranked genes varied between 20 and 300 for GTEx Skin dataset with step 5 [2].

6.1.3 Performance in presence of batch effects

One of the common real world challenges of Federated Learning applications is batch effect which refers to unwanted technical variation in gene expression data that arises from sources other than the biological differences of interest, such as differences in sample processing, RNA extraction, labeling, hybridization, and scanning. Batch effects can obscure true biological signals and generate spurious correlations between gene expression and sample attributes, leading to biased and unreliable results in downstream analyses. Therefore, batch effect correction is a critical step in the analysis of gene expression data, especially when integrating data from multiple experiments or platforms. Batch effect correction methods should be carefully applied since they could further introduce or amplify undesired effects in the data. Therefore, the quality assessment of the data integration process is crucial.

To demonstrate the robustness of Flimma towards experiential batch effects, it is applied on three additional publicly available breast cancer cohorts from GEO: GSE129508 [7], GSE149276 [8], and GSE58135 [9]. These datasets were independently collected and sequenced at three different laboratories and subjected to various experimental biases related to sample preparation, library construction, and sequencing platform (Additional file 7: Table S6). However, it is assumed that collaborating partners can agree to use the same quantification pipeline and therefore obtained uniformly (*in silico*) preprocessed raw read counts from ARCHS4 [10].

In contrast to "TCGA-BRCA", cohort-specific batch effects in the GEO datasets were much more pronounced. Principal component analysis revealed that the differences between samples from different cohorts were much larger than the differences between subtypes within the same cohort

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

(Fig. 6). In this case, effective adjustment for batch effect before testing for differential expression is crucial [11]. This can be done in two ways, either via subtracting the variation explained by batch from the data or via the inclusion of additional variables accounting for batch effects to the model. Flimma implemented the second approach, as it is preferable for downstream statistical analysis [12]. Below, it is demonstrated that this approach effectively handles the batch effects in our breast cancer data sets and gives almost identical results. Several methods for batch effect correction exist, but not all of them are compatible with limma voom because the latter is computing count-based statistics. A recently published modification of the state-of-the-art batch-effect correction method ComBat [13], namely ComBat-Seq [14], is developed specifically to handle read count data. Hence, Flimma utilized the results of limma voom obtained on the centralized GEO cohort after the removal of laboratory-specific effects by ComBat-Seq as a gold standard in the following experiments.

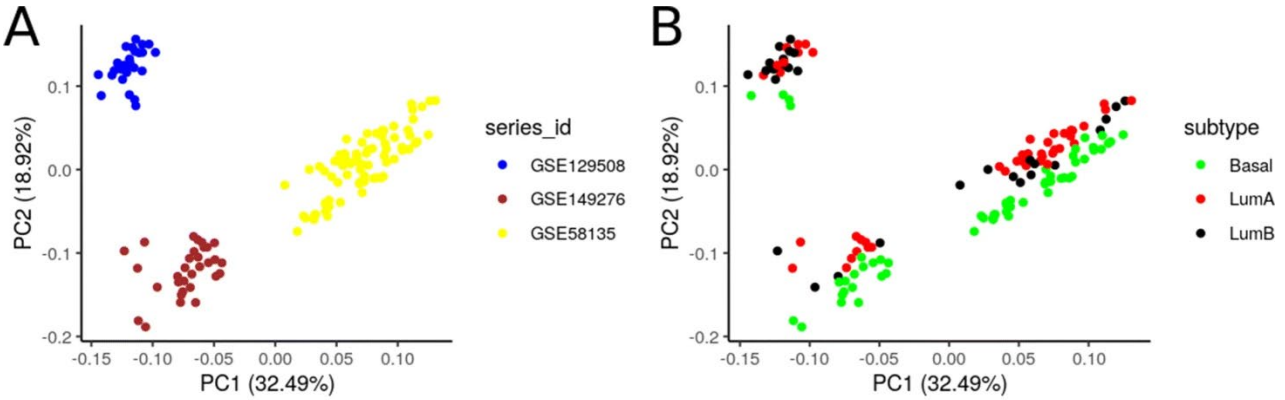

Figure 7. PCA projections computed and plotted by proBatch R package [15] of samples from three GEO cohorts (A, B) colored accordingly.[2]

Flimma models the batch effects of datasets by adding m−1 binary covariates to the linear model, where m is the number of datasets. Despite the strong batch effects in the GEO data, Flimma returned nearly the same fold-changes and BH-adjusted p-values as limma voom run on the same data after batch effect removal by ComBat-Seq (Fig. 7). Moreover, our results suggest that the approach used by Flimma gives better results than batch effect correction based on one or several first principal components.

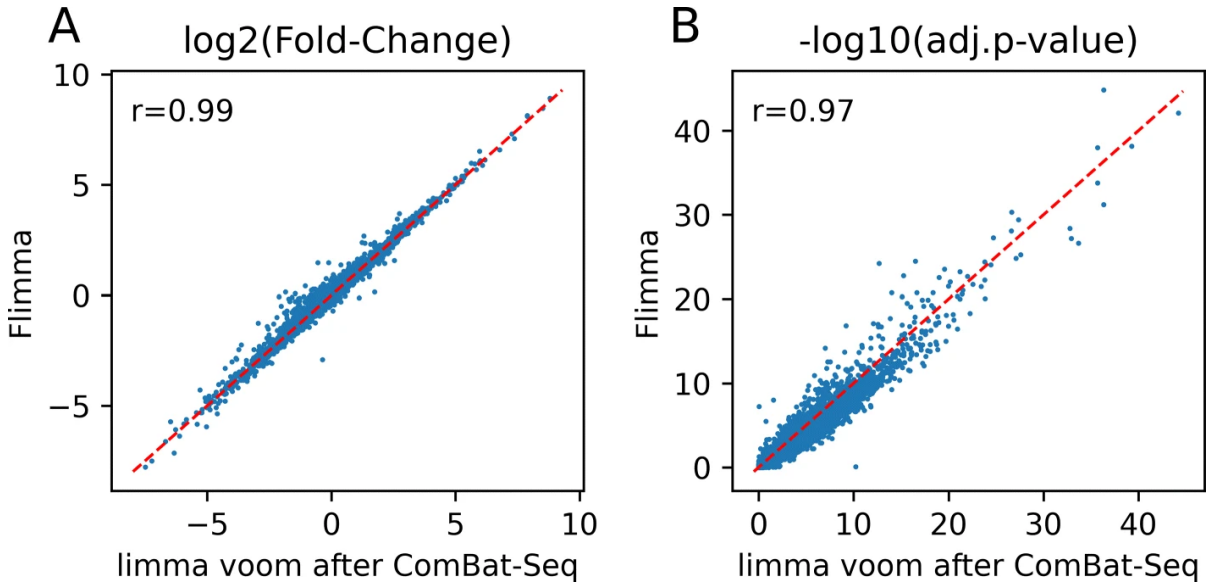

Figure 8. Comparison of the results obtained by Flimma on uncorrected GEO data with the results of limma voom after batch effect removal by ComBat-Seq [2].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

6.2 Partea

Partea performed analysis on three public benchmark datasets, commonly used in time-to-event analysis: veterans administration lung cancer Research data [39] (veterans, 137 samples), NCCTG lung cancer data [40] (lung, 168 samples). Each dataset was randomly and evenly split into 3, 5, and 10 splits to simulate different federation scenarios with different numbers of sites and sample sizes. Partea calculates survival functions for each federation scenario using hybrid approaches of FL and additive secret sharing and compares it with the central survival function estimated from the prior art lifeline.

6.2.1 Survival function

Partea compared two approaches, FL and sFL (FL and additive secret sharing), for calculating survival functions in federated scenarios. Results were compared to the central analysis using lifelines. Both FL and sFL approaches produced identical survival functions to the central analysis across different datasets and scenarios. The survival curves were presented in Figure 9. The study demonstrated that FL and sFL approaches provide equivalent results to the central analysis due to shared statistical methods.

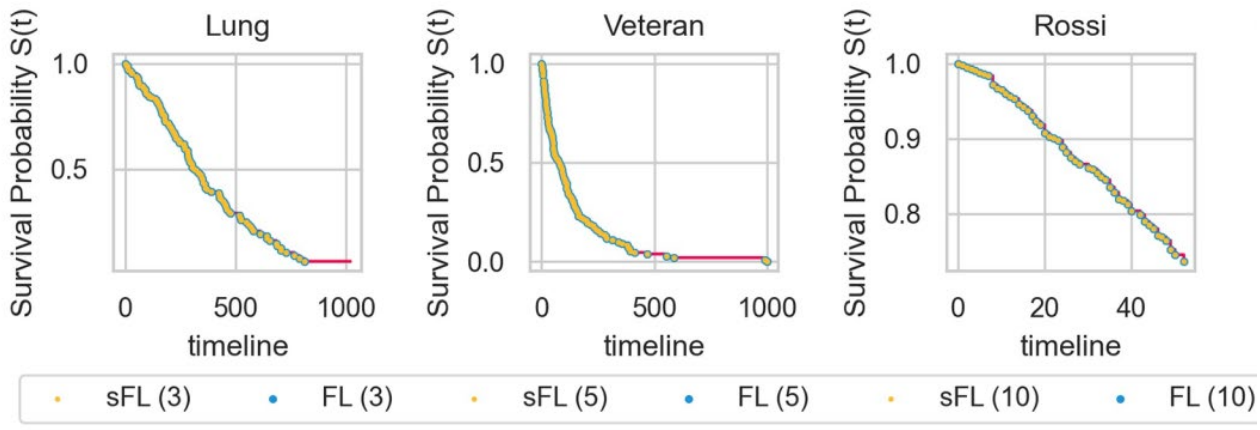

Figure 9. Evaluation of the survival function on benchmark datasets. For both the hybrid approach of FL and additive secret sharing (sFL, yellow) and the federated-only approach (FL, blue), identical survival functions are achieved compared to lifelines' Kaplan-Meier estimator (lifelines, red) for all four datasets and the various number of participants [3].

6.2.2 Differentially private survival functions

Partea [3] also incorporated differentially private (DP) survival functions and compared them to non-DP survival functions. The aim was to determine the privacy loss metric epsilon for future time-to-event analyses. The evaluation was independent of the federated computation and yielded identical results. Simulations were conducted with different epsilons (3, 2, 1, and 0.75) on each dataset (Figure 10). The log-rank test was used to compare the differentially private and non-differentially private functions. Smaller epsilon values resulted in greater differences between the DP and non-DP survival functions, particularly for smaller sample sizes. Epsilons of 3 and 2 generally showed no significant differences, while epsilon values of 1 and 0.75 occasionally led to significant differences. The findings were consistent with previous research on DP. Three predefined epsilons (3, 1 and 0.75) were suggested for future use based on the analysis and sample sizes.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

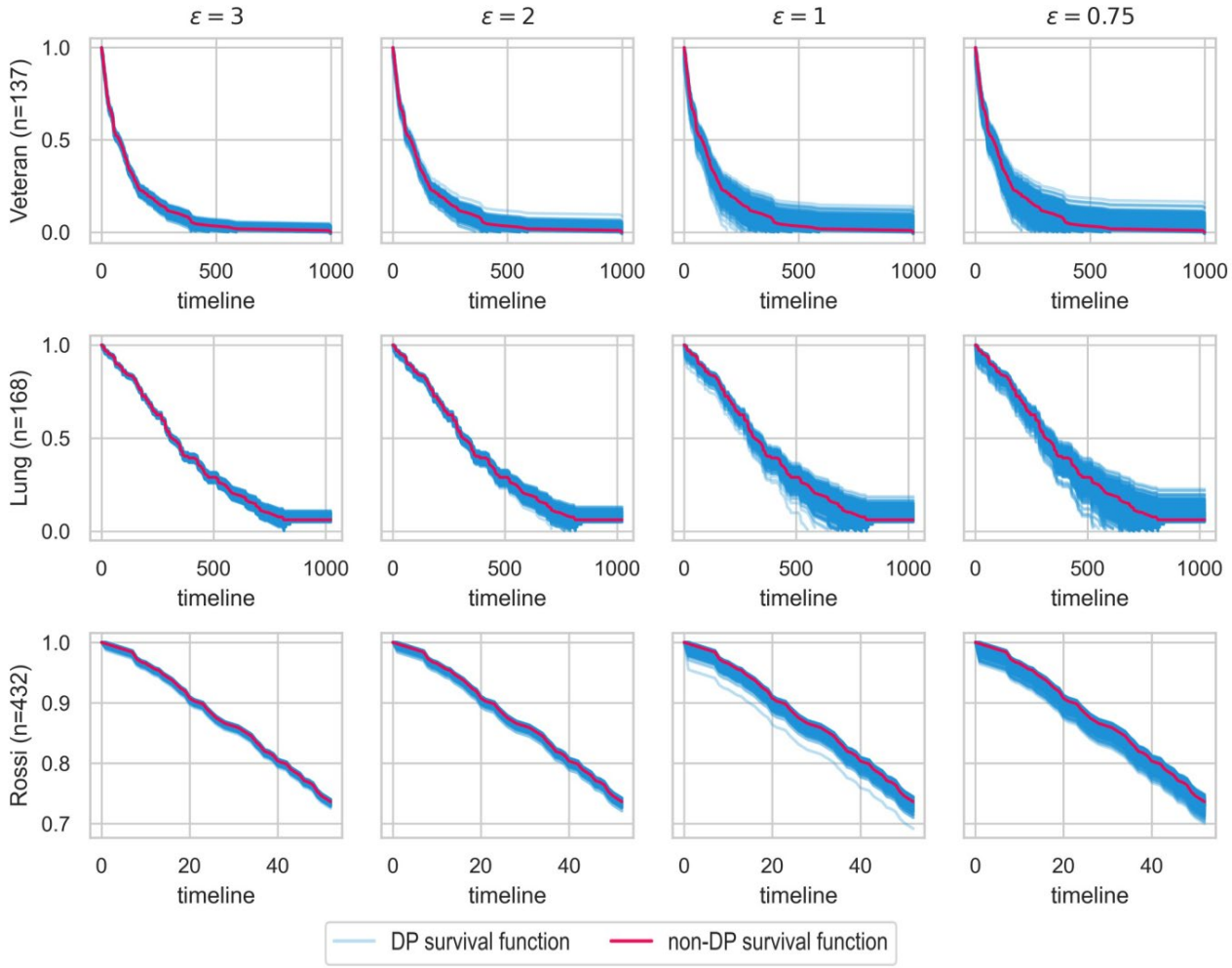

Figure 10. Comparison of DP survival functions against the non-DP baseline. The non-DP survival function (red) is used as a baseline against 1000 runs of DP survival functions for different epsilons and datasets. The resulting DP survival functions (blue) become noisier with decreasing epsilon. Note that the influence of the noise increases with decreasing sample size [3].

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

6.2.3 Cox proportional hazards model

Partea used the Cox proportional hazards model to evaluate a federated scenario. The researchers compared the resulting logarithmized hazard ratio (HR) and its 95% confidence interval (CI) for different covariates. Both the federated-only approach and the hybrid approach yielded nearly identical hazard ratios and corresponding CIs across multiple datasets and participant numbers. A comprehensive breakdown of the comparison for each covariate and dataset can be found in Fig. 11.

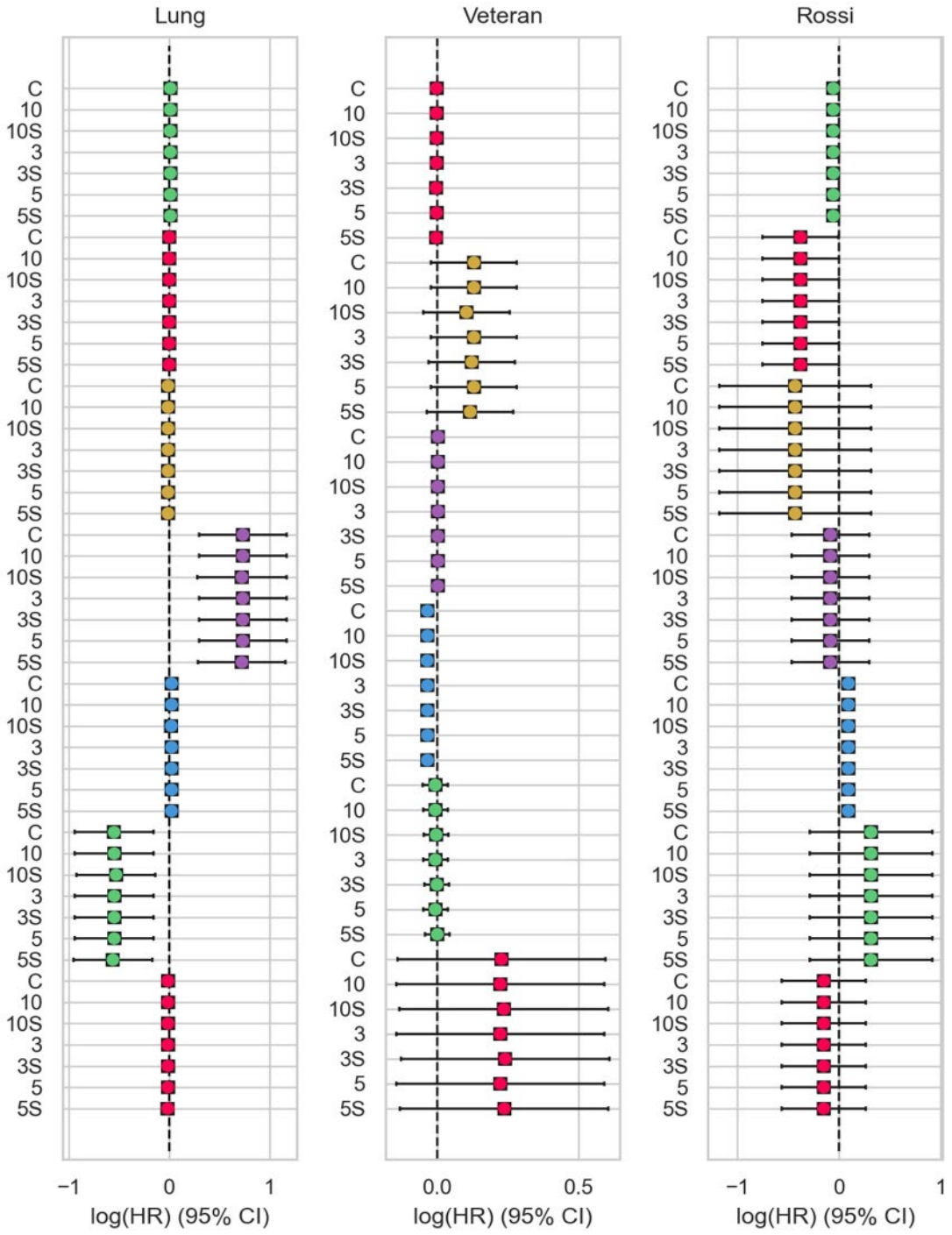

Figure 11. Evaluation of the Cox proportional hazards model on benchmark datasets. For each dataset, Partea compared the logarithmized hazard ratio and corresponding 95% CI of our algorithms for 3, 5, and 10 clients with the results of the centralized lifelines model. For all covariates

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

(distinguished by colors), the federated-only (3, 5, 10) and hybrid approach (S3, S5, S10) resulted in almost identical results compared to the centralized calculation using lifelines [3].

6.3 sPLINK

sPLINK [4] is compared with existing meta-analysis tools (PLINK, METAL, and GWAMA) using COPDGene dataset. COPDGene had 5,343 samples with an equal distribution of cases and controls for COPD analysis. sPLINK's performance was evaluated in terms of single nucleotide polymorphisms (SNP) analysis, considering confounding factors such as sex, age, smoking status, and pack years of smoking in COPDGene. The comparison aimed to assess sPLINK's effectiveness in genetic association studies, providing insights into its performance compared to other tools.

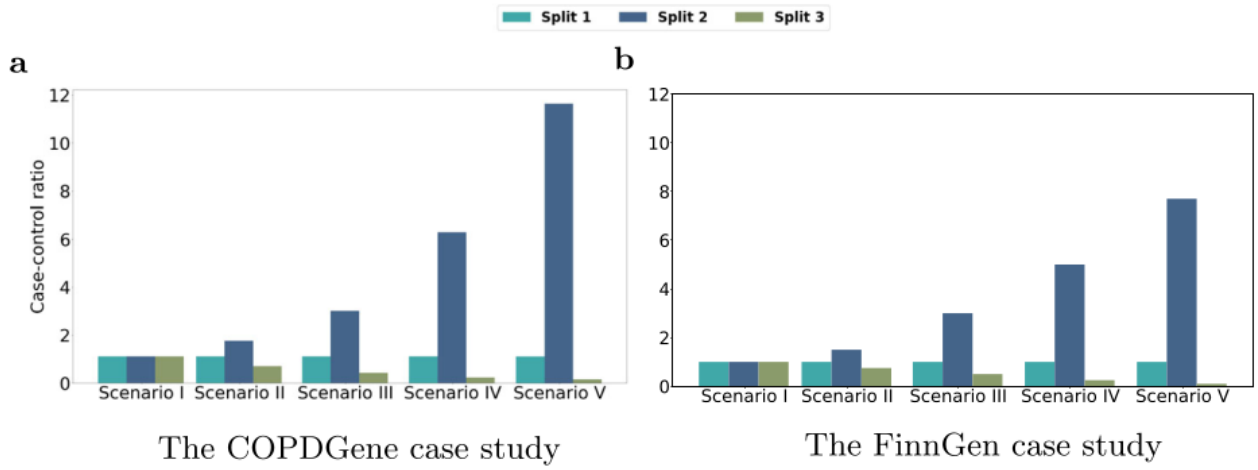

Figure 12. Scenario I-V: The case-control ratio is the same for all splits in the balanced scenario (I) while the splits have different case-control ratios in the imbalanced scenarios (II–V). All three splits have the same sample size in the COPDGene dataset as well as the balanced scenario in the FinnGen dataset. For the imbalanced scenarios in the FinnGen dataset, the splits have different sample sizes [4].

The study simulated cross-study heterogeneity on the COPDGene dataset using six different scenarios (Figure 12: scenarios 1-5, and Figure 13: scenario 6). These scenarios involve varying degrees of balance or imbalance in case-control ratios and distribution of confounding factors. The dataset was partitioned into three splits, and summary statistics were obtained for each split to conduct meta-analyses using sPLINK and other meta-analysis tools [4].

Results showed that sPLINK had a high correlation (close to 1.0) of p-values with the aggregated analysis for all scenarios, indicating consistent results regardless of phenotype or confounding factor distributions. In contrast, the correlation coefficient for other meta-analysis tools decreased with increasing imbalance or heterogeneity, suggesting reduced accuracy.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

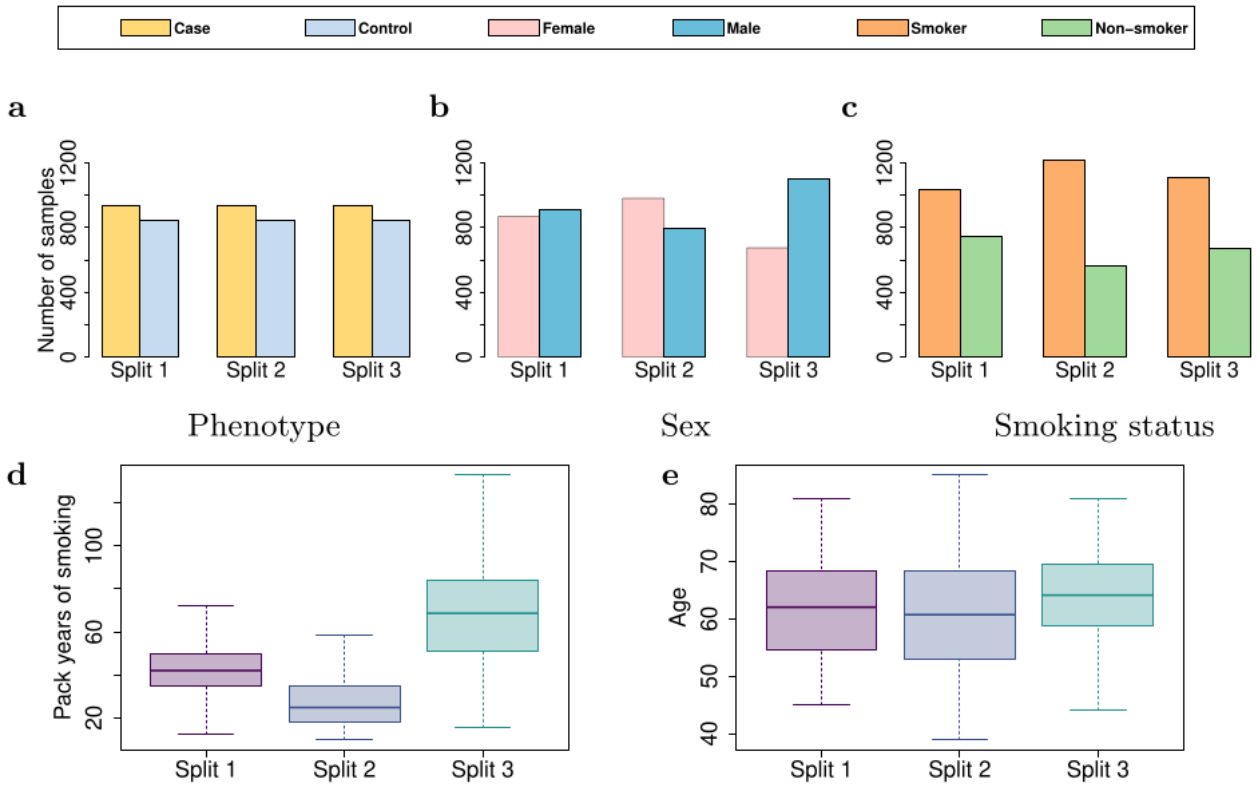

Figure 13. Scenario VI (Heterogeneous Confounding Factor) for the COPDGene case study: The phenotype distribution is the same and balanced; the values of smoking status and age are homogeneously distributed; the distribution of sex and pack years of smoking are slightly and highly heterogeneous across the splits, respectively [4].

sPLINK correctly identified all significant SNPs in all scenarios, while other meta-analysis tools missed significant SNPs, especially in highly imbalanced scenarios. False positives were minimal, with sPLINK having no false positives in any scenario, and other meta-analysis tools introducing zero or one false positive, depending on the scenario. Overall, sPLINK demonstrated robust performance in capturing significant SNPs and maintaining accurate results even in scenarios of cross-study heterogeneity, outperforming other meta-analysis tools.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

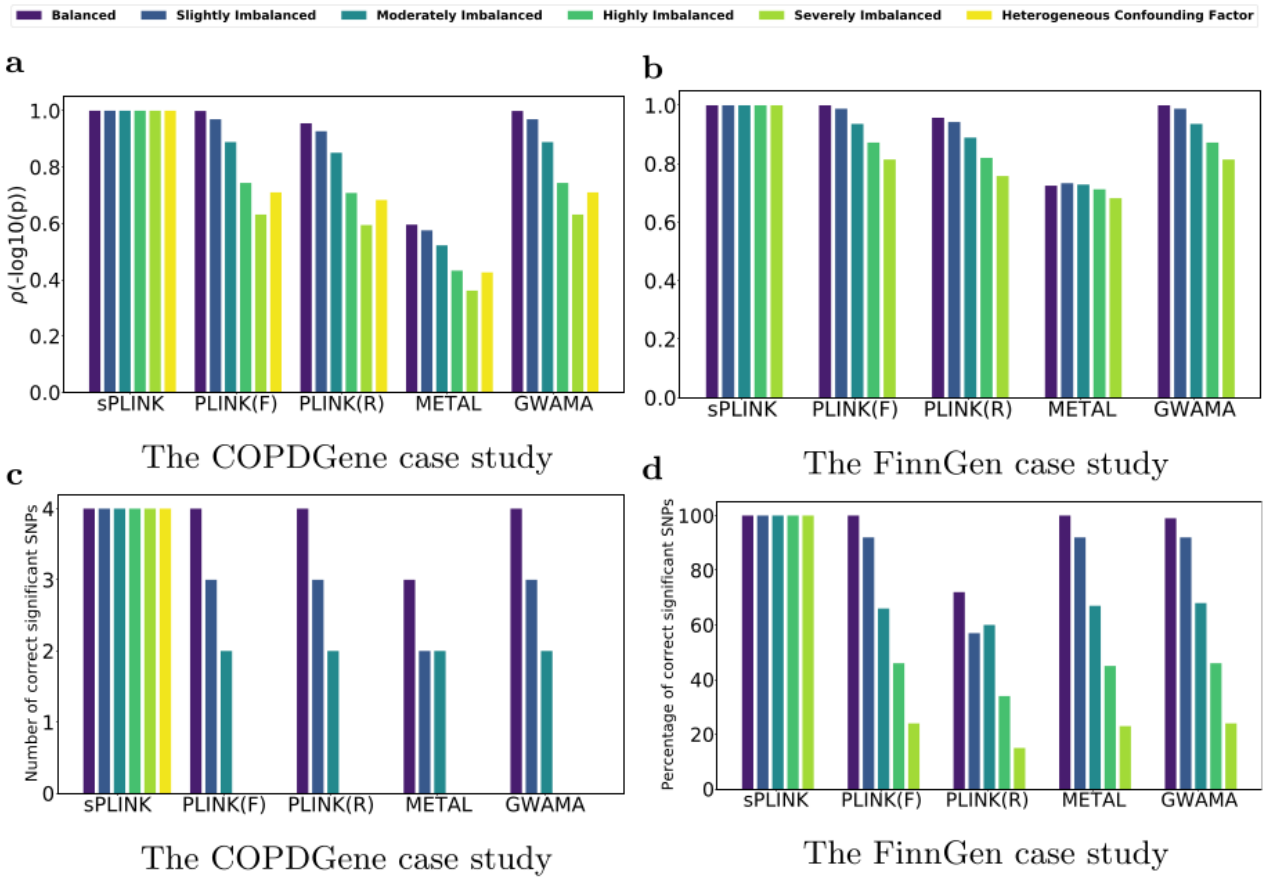

Figure 14. The Pearson correlation coefficient (ρ) of -log10 (p-value) between each tool and aggregated analysis (a, b) and the number (c) and the percentage (d) of SNPs correctly identified as significant (true positives) by each tool. F and R stand for fixed-effect and random-effect, respectively [4].

7 Open issues

No open issues.

8 Conclusion

In this deliverable we demonstrate how federated learning and in specific FeatureCloud works on publicly available dataset. In a series of peer-reviewed publications on various domains, approaches and datasets, according to the feedback from multiple reviewers of peer-reviewed journals, FeatureCloud acheives comparable results with centralized training while respecting privacy concerns by utilizing privacy enhancing technologies like Secure Multiparty Computation (SMPC) and differential privacy. FeatureCloud conducted such experiments while providing a platform and app store that presents publicly available applications to reproduce the results or extending the experiments on new fields by the community. Overall, Three approaches, Flimma (on Veteran(US Veterans' Administration lung cancer study data), Lung(NCCTG lung cancer data), and Rossi(Criminal recidivism) datasets), Partea (on COPDGene chronic obstructive pulmonary disease, and sPLINK on the "TCGA-BRCA" dataset, yield encouraging results on different criteria.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.

9 References

[1] Matschinske, J., Späth, J., Nasirigerdeh, R., Torkzadehmahani, R., Hartebrodt, A., Orbán, B., ... & Baumbach, J. (2021). The FeatureCloud AI Store for federated learning in biomedicine and beyond. arXiv preprint arXiv:2105.05734.

[2] Zolotareva, O., Nasirigerdeh, R., Matschinske, J., Torkzadehmahani, R., Bakhtiari, M., Frisch, T., ... & Baumbach, J. (2021). Flimma: a federated and privacy-aware tool for differential gene expression analysis. Genome biology, 22(1), 1-26.

[3] Späth, J., Matschinske, J., Kamanu, F. K., Murphy, S. A., Zolotareva, O., Bakhtiari, M., ... & Baumbach, J. (2022). Privacy-aware multi-institutional time-to-event studies. PLOS Digital Health, 1(9), e0000101.

[4] Nasirigerdeh, R., Torkzadehmahani, R., Matschinske, J., Frisch, T., List, M., Späth, J., ... & Baumbach, J. (2022). sPLINK: a hybrid federated tool as a robust alternative to meta-analysis in genome-wide association studies. Genome Biology, 23(1), 1-24.

[5] Nasirigerdeh, R., Torkzadehmahani, R., Matschinske, J., Baumbach, J., Rueckert, D., & Kaissis, G. (2021). HyFed: A hybrid federated framework for privacy-preserving machine learning. arXiv preprint arXiv:2105.10545.

[6] Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). Review The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemporary Oncology/Współczesna Onkologia, 2015(1), 68-77.

[7] Ligibel, J. A., Dillon, D., Giobbie-Hurder, A., McTiernan, A., Frank, E., Cornwell, M., ... & Irwin, M. L. (2019). Impact of a Pre-Operative Exercise Intervention on Breast Cancer Proliferation and Gene Expression: Results from the Pre-Operative Health and Body (PreHAB) StudyExercise Window Trial in Newly Diagnosed Breast Cancer. *Clinical Cancer Research*, *25*(17), 5398-5406.

[8] Park, S., Lee, E., Park, S., Lee, S., Nam, S. J., Kim, S. W., ... & Park, Y. H. (2020). Clinical Characteristics and Exploratory Genomic Analyses of Germline BRCA1 or BRCA2 Mutations in Breast CancerComprehensive Genomic Profile of gBRCA1/2 Breast Cancer. Molecular Cancer Research, 18(9), 1315-1325.

[9] Varley, K. E., Gertz, J., Roberts, B. S., Davis, N. S., Bowling, K. M., Kirby, M. K., ... & Myers, R. M. (2014). Recurrent read-through fusion transcripts in breast cancer. Breast cancer research and treatment, 146, 287-297.

[10] Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., ... & Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. Nature communications, 9(1), 1366.

[11] Leek, J. T., Scharpf, R. B., Bravo, H. C., Simcha, D., Langmead, B., Johnson, W. E., ... & Irizarry, R. A. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, *11*(10), 733-739.

[12] Nygaard, V., Rødland, E. A., & Hovig, E. (2016). Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostatistics*, *17*(1), 29-39.

[13] Adjusting batch effects in microarray expression data using empirical Bayes methods.

[14] Zhang, Y., Parmigiani, G., & Johnson, W. E. (2020). ComBat-seq: batch effect adjustment for RNA-seq count data. NAR genomics and bioinformatics, 2(3), lqaa078.

[15] Čuklina, J., Lee, C. H., Williams, E. G., Sajic, T., Collins, B. C., Rodríguez Martínez, M., Pedrioli, P. G. (2021). Diagnostics and correction of batch effects in large-scale proteomic studies: A tutorial. Molecular systems biology, 17(8), e10240.

[16] Nasirigerdeh, R., Torkzadehmahani, R., Matschinske, J., Baumbach, J., Rueckert, D., & Kaissis, G. (2021). HyFed: A hybrid federated framework for privacy-preserving machine learning. arXiv preprint arXiv:2105.10545.

[17] GTEx Consortium. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science, 369(6509), 1318-1330.

[18] Therneau, T., & Lumley, T. (2013). R survival package. R Core Team.

[19] Davidson-Pilon, C. (2019). lifelines: survival analysis in Python. Journal of Open Source Software, 4(40), 1317.

[20] COPDGene. http://www.copdgene.org/. Accessed 30 Nov 2021.

[21] FinnGen Consortium. (2021). FinnGen documentation of R4 Release.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078.