Distributed Analytics for Research in Hospitals (DARAH): Federated Analysis with Differential Privacy in a Real-World Oncology Study

Théo Riffel, Perrine Créquit, Maëlle Baillet, Jason Paumier, Yasmine Marfoq, Ronan Sy, Olivier Girardot, Thierry Chanet, Louise Bayssat, Julien Mazières, Vincent Vuiblet, Julien Ancel, Maxime Dewolf, François Margraff, Camille Bachot, Jacek Chmiel

Submitted to: JMIR Medical Informatics
on: April 19, 2024

Table of Contents

Distributed Analytics for Research in Hospitals (DARAH): Federated Analysis with Differential Privacy in a Real-World Oncology Study

Théo Riffel[1] PhD; Perrine Créquit[2] MD; Maëlle Baillet[1] MD; Jason Paumier[1] MSc; Yasmine Marfoq[1] MSc; Ronan Sy[1] MSc; Olivier Girardot[1] MSc; Thierry Chanet[1] MSc; Louise Bayssat[1] MSc; Julien Mazières[3, 4, 5] MD, PhD; Vincent Vuiblet[6] MD; Julien Ancel[7, 8] MD; Maxime Dewolf[7] MD; François Margraff[9] MSc; Camille Bachot[9] MSc; Jacek Chmiel[9] MSc

[1]Arkhn Paris FR
[2]Hôpital Foch Suresnes FR
[3]INSERM Toulouse FR
[4]Centre de Recherches en Cancérologie Oncopole Toulouse FR
[5]Centre Hospitalier Universitaire de Toulouse Service de Pneumologie Toulouse FR
[6]Centre Hospitalier Universitaire de Reims Service de Néphrologie Reims FR
[7]Centre Hospitalier Universitaire de Reims Service de Pneumologie Reims FR
[8]INSERM Reims FR
[9]Roche SAS Boulogne FR

Corresponding Author:
Théo Riffel PhD
Arkhn
9, rue d'Alexandrie
Paris
FR

Abstract

Background: Federated analytics in healthcare allows researchers to perform statistical queries on remote data sets without access to the raw data. This method arose from the need to perform statistical analysis on larger datasets collected at multiple healthcare centers while avoiding regulatory, governance, and privacy issues that might arise if raw data were collected at a central location outside the healthcare centers. Despite some pioneering work, federated analytics is still not widely used on real-world data, and to our knowledge, no real-world study has yet combined it with other privacy-enhancing techniques such as differential privacy.federated analysis, differential privacy, real-world oncology study, non-small cell lung cancer, COVID-19federated analysis, differential privacy, real-world oncology study, non-small cell lung cancer, COVID-19

Objective: The first objective of this study was to deploy a federated architecture in a real-world setting. The oncology study used for this deployment compared the medical healthcare management of patients with metastatic non-small cell lung cancer before and during/after the 1st wave of COVID-19. The second goal was to test differential privacy in this real-world scenario to assess its practicality and utility as a privacy enhancing technology.

Methods: A federated architecture platform was set up in the Toulouse, Reims and Foch centers. After harmonization of the data in each center, statistical analyses were performed using DataSHIELD, a federated analysis R library and a new open source differential privacy DataSHIELD package was implemented: dsPrivacy.

Results: 50 patients were enrolled in the Toulouse and Reims centers and 49 in the Foch center. We have shown that DataSHIELD is a practical tool to efficiently conduct our study across all 3 centers without exposing data on a central node, once sufficient setup has been made to configure a secure network between hospitals. All planned aggregated results were successfully generated. We also observed that differential privacy can be implemented in practice with promising trade-offs between privacy and accuracy, and we built a library that will prove useful for future work.

Conclusions: The federated architecture platform enabled a multicenter study to be conducted on real-world oncology data with strong privacy guarantees thanks to differential privacy.

(JMIR Preprints 19/04/2024:59685)
DOI: https://doi.org/10.2196/preprints.59685

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Original Paper

Distributed Analytics for Research in Hospitals (DARAH): Federated Analysis with Differential Privacy in a Real-World Oncology Study

Authors
Théo Ryffel, PhD, Arkhn, Paris, France
Perrine Créquit, MD, Hôpital Foch, Paris, France
Maëlle Baillet, MD, Arkhn, Paris, France
Jason Paumier, MSc, Arkhn, Paris, France
Yasmine Marfoq, MSc, Arkhn, Paris, France
Ronan Sy, MSc, Arkhn, Paris, France
Olivier Girardot, MSc, Arkhn, Paris, France
Thierry Chanet, MSc, Arkhn, Paris, France
Louise Bayssat, MSc, Arkhn, Paris, France
Julien Mazières, MD PhD, Service de Pneumologie, CHU Toulouse, INSERM, CRCT Oncopole, Toulouse, France
Vincent Vuiblet, MD, Service de Néphrologie, CHU de Reims, Reims, France
Julien Ancel, MD, Service de Pneumologie, CHU de Reims, INSERM, Reims, France
Maxime Dewolf, MD, Service de Pneumologie, CHU de Reims, Reims, France
François Margraff, MSc, Roche SAS, Boulogne, France
Camille Bachot, MSc, Roche SAS, Boulogne, France
Jacek Chmiel, MSc, Avenga, Warsaw, Poland

Corresponding author
Théo Ryffel
9 rue d'Alexandrie, 75002 Paris, France
theo.ryffel@arkhn.com
+33 6 45 21 57 53

Keywords: federated analysis; differential privacy; real-world oncology study; non-small cell lung cancer; COVID-19

Abstract:

Background: Federated analytics in healthcare allows researchers to perform statistical queries on remote data sets without access to the raw data. This method arose from the need

to perform statistical analysis on larger datasets collected at multiple healthcare centers while avoiding regulatory, governance, and privacy issues that might arise if raw data were collected at a central location outside the healthcare centers. Despite some pioneering work, federated analytics is still not widely used on real-world data, and to our knowledge, no real-world study has yet combined it with other privacy-enhancing techniques such as differential privacy.

Introduction: The first objective of this study was to deploy a federated architecture in a real-world setting. The oncology study used for this deployment compared the medical healthcare management of patients with metastatic non-small cell lung cancer before and during/after the 1st wave of COVID-19. The second goal was to test differential privacy in this real-world scenario to assess its practicality and utility as a privacy enhancing technology.

Methods: A federated architecture platform was set up in the Toulouse, Reims and Foch centers. After harmonization of the data in each center, statistical analyses were performed using DataSHIELD, a federated analysis R library and a new open source differential privacy DataSHIELD package was implemented: dsPrivacy.

Results: 50 patients were enrolled in the Toulouse and Reims centers and 49 in the Foch center. We have shown that DataSHIELD is a practical tool to efficiently conduct our study across all 3 centers without exposing data on a central node, once sufficient setup has been made to configure a secure network between hospitals. All planned aggregated results were successfully generated. We also observed that differential privacy can be implemented in practice with promising trade-offs between privacy and accuracy, and we built a library that will prove useful for future work.

Conclusion: The federated architecture platform enabled a multicenter study to be conducted on real-world oncology data with strong privacy guarantees thanks to differential privacy.

Introduction

Federated analytics (FA) [1] allow researchers to perform statistical queries on remote datasets without accessing raw data. This method has emerged from the need of conducting statistical analyses on larger datasets originating from multiple healthcare centers, while avoiding regulation, governance and privacy related issues that could occur if the raw data was gathered in a central location external to the healthcare centers.

DataSHIELD [2] is one of the pioneering open-source software solutions for bioscience collaboration using federated analysis. It has been used in several projects including the EU Child Cohort Network [3] and German MIRACUM consortium [4]. However, even if it has been identified as a next step in some oncology studies [5], to the best of our knowledge it has not been used on real world oncology data. One possible reason for this can be that data harmonization is actually the sticking point, since it proves to be challenging on real world data when not already done ahead of time.

While FA ensures that sensitive data is never directly exposed, results from statistical queries can still leak some information from individuals. Indeed, several privacy attacks [6, 7] have been proposed to exploit common statistical analysis results and disclose private information. To mitigate these attacks, differential privacy can be used in combination with FA to provide stronger privacy guarantees. Differential privacy (DP) [8, 9] is a method for computing statistical analyses on a sensitive dataset in such a way that the results do not compromise the privacy of the initial raw data. It proves particularly appropriate for studies on patient health data which are highly sensitive, but as far as we know, no real world data study has already combined FA and privacy-enhancing techniques such as differential privacy.

The goal of this study, called Distributed Analytics for Research in Hospitals (DARAH), is to evaluate the operational deployment of a federated architecture in the context of a real-world oncology study. Another objective is to analyze the impact of differential privacy in a federated analysis on the meaningfulness of the study results.

Methods

Use of DataSHIELD

The use of DataSHIELD was recommended by Roche based on previous experiences with privacy-preserving analytics of medical data. Key points assessed were the maturity of the solution, especially in terms of security, stability and community support.
DataSHIELD showed more maturity than any other open-source framework available in terms of privacy, with default privacy disclosure policies, account management and permissions management (critical for federated setups with limited trust between parties) which are handled locally by each center. It is also packaged with a large set of ready-to-use federated statistical functions. In addition, DataSHIELD proved not only very stable but also very flexible, with built-in extensions mechanisms from custom community libraries.
At the time of starting the project, in Roche's Federated Open Science team view, it was the only open source tool for federated analytics that addressed the threat model of "honest but curious" data scientists, with strong and highly configurable built-in privacy filters and safeguards against abusive centralized authorization management or arbitrary code execution. In addition, a key advantage of DataSHIELD is the set of ready to use, fully matured packages of federated statistics functions written in R (dsBase) package, which lower the entry barrier for data scientists to analyze medical data. The DataSHIELD community was assessed as relatively large and responsive, including official forums and direct messages from DataSHIELD developers, which in turn minimized the risk of using open source solutions without sufficient support. It was already known that DataSHIELD dsBase functions were lacking a built-in differential privacy mechanism, but it was considered feasible to extend the capabilities of DataSHIELD with custom-built functions and third-party differential privacy libraries.
DataSHIELD was kept in its default privacy disclosure mode, which is a bit more relaxed than strict privacy models, and all the privacy filters were also kept using their default settings, without lowering or increasing their privacy enforcement strength. Default setting also included keeping TLS (SSL) enforcement for https connections despite the ability of version 6.2 to override those checks and everything happening in the isolated VPN network. It required local management of certificates to avoid blocking connections to external nodes. Hospital nodes use locally managed users, which is both DataSHIELD's default and proper strategy for authentication and user management in a federated network.

Use of Differential Privacy

Differential privacy provides a measurement of the privacy risk associated with publishing each particular result on a sensitive dataset, by measuring the maximum leakage that each result can cause about the individuals' data. In practice, it often works through the addition of a controlled amount of statistical noise to obscure the contributions to the result from each individual in the dataset.

Several open-source libraries exist that implement differential privacy, most of which are written in

Python or expose a Python interface. Popular ones include Google's differential privacy library [10], PyDP [11] from the OpenMined community and OpenDP [12]. In R, the main open-source library is the diffpriv library [13] that implements some differential privacy mechanisms but not a real suite of functions ready to use, and which is not actively supported with last contributions going back to 2017. Moreover, none of these tools were directly compatible with the DataSHIELD framework, which requires a specific interface in R. As a result, an new open-source a differential privacy R package for DataSHIELD has been implemented: **dsPrivacy**[1].

DsPrivacy implements common statistical operations such as mean or standard deviation with differential privacy and can be directly integrated into DataSHIELD. More specifically, pure differential privacy has been implemented, which means that the statistical noise added to the result is drawn from a Laplacian distribution. Privacy is therefore defined with a single parameter ε (which corresponds to the inverse of the noise variance) and it is called the *privacy budget*. The privacy budget ε controls the amount of noise injected in the computation and hence the privacy of the analysis: the smaller the privacy budget ε the higher the noise, which means better privacy guarantees since the private information is better covered by the noise.

Differential privacy induces a trade-off between privacy and accuracy of the analysis [29]: as privacy is ensured with noise, better privacy means a lower signal-to-noise ratio and less meaningful results. Conversely, when ε is high, the noise is reduced and privacy degrades. Choosing the right value for ε from a privacy standpoint is quite controversial [14] and is highly dependent on the context and the study considered.

DsPrivacy in the context of FA works by adding Laplacian noise locally at each center on the results computed on the local datasets before results are aggregated on a central node. This is referred to as local differential privacy, as opposed to central differential privacy, where the noise is added after the aggregation part on the central node. This paradigm is detailed in the discussion.

1 Available at https://github.com/arkhn/dsPrivacy

Use case: Differences in the patient drug exposure of non-small cell lung cancer (NSCLC) patients before and after the first wave of COVID-19

Context

In order to test the implementation of a federated platform with differential privacy, a study was conducted using real-world oncology data, with the aim of analyzing the differences in the patient drug exposure of non-small cell lung cancer patients before (BW) and after (AW) the first wave of COVID-19. Specifically, the analyses were based on the first line of treatment, including its duration and disease progression at 24 months. The study was carried out in three hospitals: Foch Hospital, Toulouse University Hospital Center and Reims University Hospital Center.

Patient Selection

The inclusion criteria were adult patients with metastatic NSCLC treated by chemotherapy and/or immunotherapy and/or antiangiogenic at Foch, Toulouse and Reims centers between March 2019 and March 2021. The list of considered treatments is available in Appendix A.

The exclusion criteria were patients objecting to the reuse of their data, protected adult patients (patients under curatorship, tutorship or advisership), patients undergoing a clinical trial, patients whose follow-up did not start in one of these centers, patients who are not immediately metastatic and patients with an EGFR mutation, ALK translocation or ROS1 mutation.

For this study, the objective was to have 25 patients per period and per site (i.e. 50 patients per center). Initially, patients were pre-selected using the inclusion criteria present in the chemotherapy prescription software CHIMIO, namely the date and type of treatment. Since the other criteria for inclusion and exclusion were not available in CHIMIO, physicians in each center manually assessed whether the patients could be included in the study.

From an ethical and regulatory standpoint, this study was conducted in accordance with the MR-004

reference methodology.

Variables

In this study, the variables of interest were the duration of the first line of treatment and the disease progression at 24 months following or during the first line of treatment. The start of a second line of treatment or death following (before the start of a second line) or during the first line of treatment were the proxies used for disease progression at 24 months. These events had to have occurred during the observation period, which was 2 years after the patient's inclusion in the study.

The explanatory variable was the patient's inclusion period: before the first wave of COVID-19 (BW): March 1, 2019 to March 1, 2020 or after the first wave of COVID-19 (AW) : March 2, 2020 to March 31, 2021.

Some potential confounders were also studied: age, gender, BMI, blood creatinine level and type of treatment.

All these variables were extracted from the CHIMIO software and are listed in Appendix B.

Data harmonization

Federated analysis requires data from different centers to be harmonized. We use FHIR as our oncology data standard [15, 16], and the data model is available in Appendix C. To facilitate the analyses, a tabularized version of the FHIR standard has been produced and the correspondence between the variables of interest used in the datasets and the FHIR attributes are presented in Appendix B. This data model has been used in each center to create harmonized local clinical datasets.

Statistical definition without differential privacy

First, univariate and bivariate descriptive statistics were performed for the variables of interest, the explanatory variables and the confounding factors. For the qualitative variables, percentages per category were calculated. For the quantitative variables, the mean but also the 5th, 25th, 50th, 75th and 95th percentiles were computed. The minimums and maximums were not computed because these values are considered disclosive (e.g. the presence of an outlier) in DataSHIELD standard configuration.

Then, to ensure the absence of confounding factors, covariates for which there was a difference between the two groups (BW and AW) in bivariate analysis with a $P<.20$ were included in the multivariate models. A Student's t test was used for quantitative variables. A Chi-2 test was used for qualitative variables with at least 5 elements in each class. Otherwise, a Fisher test was used.

Finally, to investigate a difference in management between patients in the first and second waves, a linear regression was performed. The variable of interest was the duration of first-line treatment at the metastatic stage and the explanatory variable was the period (before the first wave / after the first wave). The duration of treatment was defined as the period from day 1 to the last day of the first line. A multivariate survival analysis (Cox model) was also performed in which the event of interest was the disease progression at 24 months. The explanatory variable was also the period (before the first wave / after the first wave). Potential confounders identified in the previous step were included in both multivariate analyses. A difference was considered significant if the confidence interval did not contain 1 for linear regression and survival analysis.

Statistical definition with differential privacy

The univariate and bivariate statistical analyses presented in above were also performed with differential privacy. The global strategy used to implement differential privacy has been discussed in section Use of Differential Privacy and some details for each part of the study are detailed here.

All the univariate analyses run code from PyDP [11] that we wrapped in dsPrivacy. As adding DP preserves privacy, min and max values could be computed instead of 5th and 95th percentiles. Then bivariate analyses were carried out. For quantitative variables, an implementation of the Student test is used and from the result of this test, a *P*-value is inferred. For qualitative variables, occurrence tables are computed (under the hood, the function tableDP from DPPack [17] is used in each center and all the tables are summed) and then a Chi-2 or Fisher test is performed with the R functions chisq.test and fisher.test. Linear regression and Cox models were not implemented with differential privacy.

As presented above, setting privacy budgets is something difficult. Analyses were performed with a budget considered “good enough” for the results to be relevant. This budget should be different for different kinds of operations: for instance, quadratic operations are much more sensitive to noise than linear operations. This explains why the budget used, which is reported in the result tables below, differs between simpler queries such as mean and other univariate queries ($\varepsilon = 5.0$) and student tests which needed a higher budget ($\varepsilon = 60.0$) to show coherent results.

Results

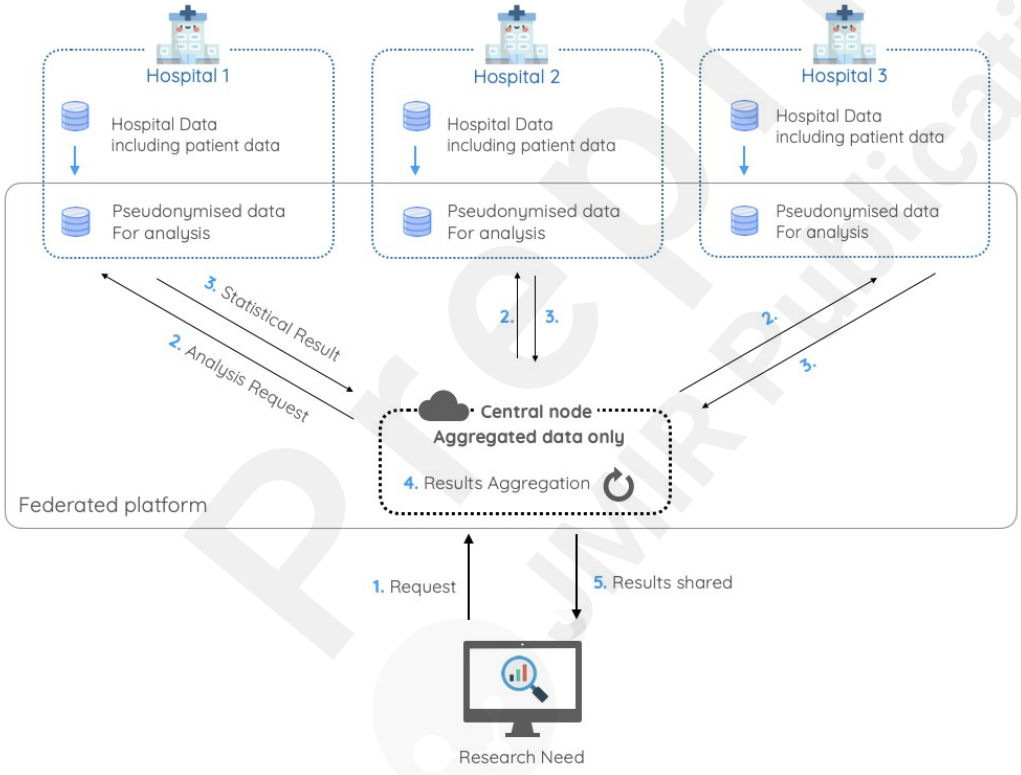
Deployment of federated architecture in the three centers

The schema of our federated architecture is given in Figure 1. Researchers connect through a VPN to a central node hosted on a cloud provider certified ISO27001. This central node coordinates the execution of the queries on the remote data assets hosted in the different hospitals and aggregate all the responses received. Each hospital exposes through the private network pseudonymized datasets which originate from their local harmonized data architecture, but which are completely isolated on a

dedicated infrastructure for security purposes. Each hospital receives and executes remote requests from the central node that match local permissions set in DataSHIELD regarding data assets, authorized functions and legitimate users. This ensures that hospitals keep full sovereignty on how their data is exploited.

One key element is to deploy a private network between all stakeholders to mitigate the risk of intrusions. This was done by using IpSec bridges between the central cloud-hosted node and the different sites and to install secure certificates that allow for secure connections inside the private network.

Figure 1: Federated architecture diagram

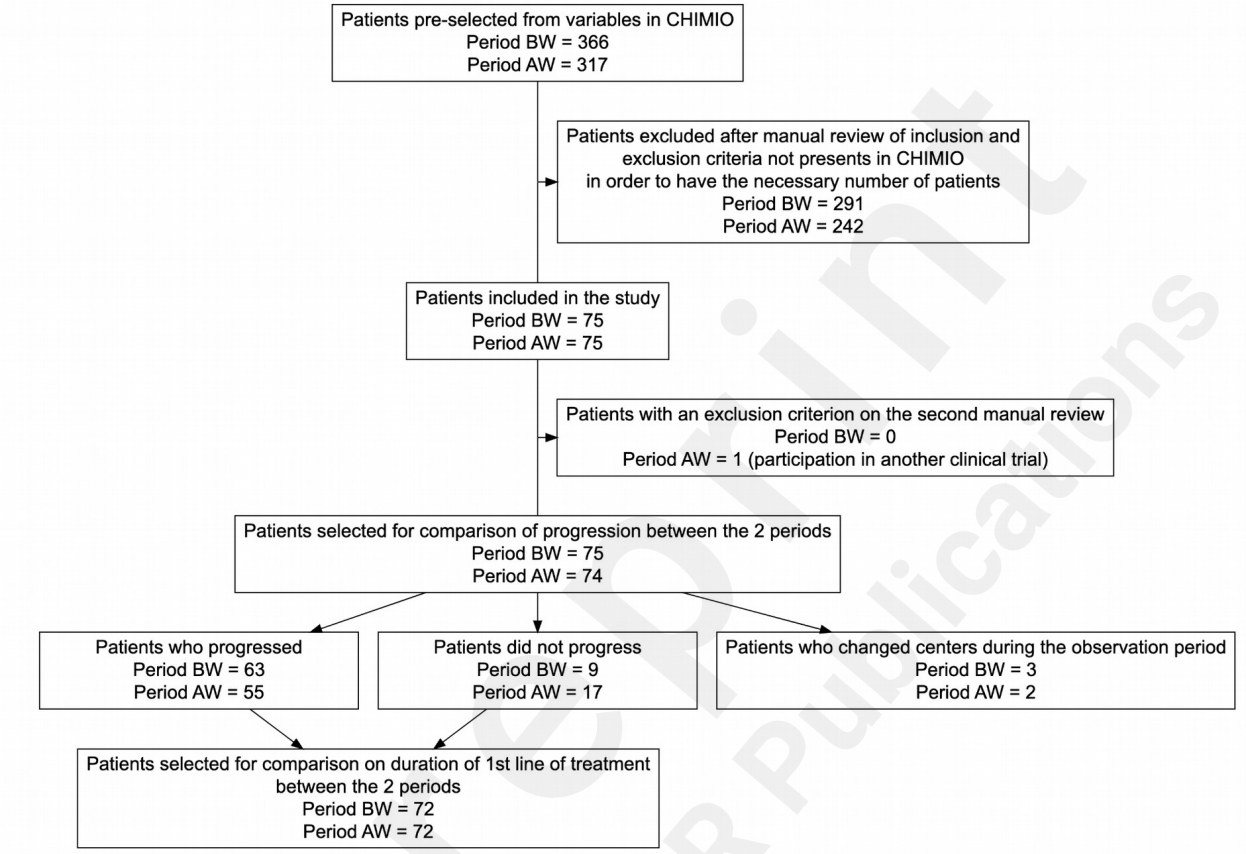


Flowchart

The flowchart is presented in Figure 2. 75 and 74 patients from BW and AW respectively were included in the analysis comparing disease progression. Of the patients in the BW period, 63 had progressed, 9 were progression-free and 3 had changed groups during the observation period and were therefore censored at the time of transfer. Of the patients in the AW period, 55 had progressed,

17 were progression-free and 2 had changed groups during the observation period and were therefore censored at the time of transfer. Thus, after exclusion of patients who changed centers, 72 patients from each period were included in the analysis of the duration of the first line of treatment.

Figure 2: Flowchart



Bivariate analysis and selection of potential confounders

Combined univariate analyses without and with differential privacy (DP) are presented in Appendix D. Combined bivariate analyses without and with DP are presented in Table 1 and 2 respectively. Direct comparison between DP and no-DP analyses across all periods (AW+BW) is available in Table 3. In analysis without and with DP, the covariates selected for multivariate analyses were the treatment type (<0.001) only.

Comparison of the progression between two periods (Cox model)

In the analysis without DP, the hazard ratio of the cox model for the AW period compared to the BW period was 0.98 [0.65;1.46]. As the confidence interval includes 1, AW was not considered significantly different from the BW period in terms of disease progression at 24 months after adjustment for a significant covariate (treatment type).

Comparison of the duration of the first line of treatment between the two periods (linear regression)

In the analysis without DP, the coefficient of linear regression for the AW period compared to the BW period was -13.84 [-103.82;76.14]. As the confidence interval includes 1, AW was not considered significantly different from the BW period in terms of first line duration after adjustment for significant covariate (treatment type).

Table 1: Bivariate analyses without DP

Variables	Patients (n=75) Period BW	Patients (n=74) Period AW	*P*
Variables of interest			
Disease Progression, n(%)			
No	12 (16.0)	19 (25.7)	.21
Yes	63 (84.0)	55 (74.3)	
Missing	0 (0.0)	0 (0.0)	
First Line Duration (days)			
Mean (sd)	179.7 (267.1)	255.2 (274.6)	.09
Median (Q25-Q75)	75.7 (14.7-190.0)	141.8 (45.1-420.8)	
Q5-Q95	0-592.4	15-792.8	
Missing, n(%)	(0.0)	(0.0)	
Covariables			
Organization, n(%)			
Toulouse University Hospital	25 (33.3)	25 (33.8)	.99
Reims University Hospital	25 (33.3)	25 (33.8)	
Foch Hospital	25 (33.3)	24 (32.4)	
Missing	0 (0.0)	0 (0.0)	
Gender, n(%)			
Female	33 (44.0)	30 (40.5)	.79
Male	42 (56.0)	44 (59.5)	
Missing	0 (0.0)	0 (0.0)	
Age, n(%)			
<55	12 (16.0)	10 (13.5)	.38
55-65	30 (40.0)	23 (31.1)	
>65	33 (44.0)	41 (55.4)	
Missing	0 (0.0)	0 (0.0)	
BMI, n(%)			
<18.5	13 (17.3)	19 (25.7)	.39
18.5-25	34 (45.3)	27 (36.5)	
>25	28 (37.3)	28 (37.8)	
Missing	0 (0.0)	0 (0.0)	
Treatment Category, n(%)			
Chemotherapy	46 (61.3)	30 (40.5)	<.001
Chemotherapy+Angiogenesis Inhibitor	6 (8.0)	0 (0.0)	
Chemotherapy+Immunotherapy	7 (9.3)	33 (44.6)	
Immunotherapy	16 (21.3)	11 (14.9)	
Missing	0 (0.0)	0 (0.0)	
Creatinemia (µmol/l)			
Mean (sd)	66.7 (20.3)	64.3 (18.9)	.47
Median (Q25-Q75)	60.0 (54.7-73.3)	63.7 (55.1-73.1)	
Q5-Q95	46.7-98.8	41.1-82.8	
Missing, n(%)	0 (0.0)	1 (1.4)	

DP: Differential Privacy; BW: Before the wave; AW: After the wave; BMI: Body Mass Index; sd: Standard Deviation

Table 2: Bivariate analyses with DP

Variables	Patients Period BW	Patients Period AW	*P*
Variables of interest			
Disease Progression, n(%)	ε = 5.0		
No	12 (16.0)	18 (24.7)	.27
Yes	63 (84.0)	55 (75.3)	
First Line Duration (days)	ε = 5.0		ε = 60.0
Mean (sd)	189.5 (231.1)	287.4 (286.3)	.051
Median	81.8	132.6	
min-max	0.3-1359.5	8.5-827.1	
Covariables			
Gender, n(%)	ε = 5.0		
Female	32 (43.2)	30 (41.1)	.92
Male	42 (56.6)	43 (58.9)	
Age, n(%)	ε = 5.0		
<55	12 (16.2)	10 (13.7)	.37
55-65	29 (39.2)	22 (30.1)	
>65	33 (44.6)	41 (56.2)	
BMI, n(%)	ε = 5.0		
<18.5	13 (17.1)	19 (25.7)	.38
18.5-25	34 (44.7)	27 (36.5)	
>25	29 (38.2)	28 (37.8)	
Treatment Category, n(%)	ε = 5.0		
Chemotherapy	46 (61.3)	31 (39.7)	.001
Chemotherapy+Angiogenesis Inhibitor	6 (8.0)	0 (0.0)	
Chemotherapy+Immunotherapy	7 (9.3)	35 (44.9)	
Immunotherapy	16 (21.3)	12 (15.4)	
Creatinemia (µmol/l)	ε = 5.0		ε = 60.0
Mean (sd)	65.4 (18.9)	61.9 (17.1)	.53
Median	64.4	65.4	
min-max	48.8-88.9	37.4-86.1	

DP: Differential Privacy; BW: Before the wave; AW: After the wave; BMI: Body Mass Index; sd: Standard Deviation; ε: privacy budget

Table 3: Comparison of bivariate analyses across all periods (AM + BW), with and without DP

Variables	Global results (BW+AW) without DP	Global results (BW+AW) with DP (ε = 5.0)
Variables of interest		
Disease Progression, n(%)		
No	31	31
Yes	118	118
First Line Duration (days)		
Mean (sd)	217.2 (269.4)	198.9 (286.0)
Median	113.5	126.4
min-max	0.5-785.6	0.1-524.8
Covariables		
Gender, n(%)		
Female	63	63
Male	86	87
Age, n(%)		

<55	22	21
55-65	53	52
>65	74	75
BMI, n(%)		
<18.5	32	32
18.5-25	61	61
>25	56	56
Treatment Category, n(%)		
Chemotherapy	76	76
Chemotherapy+Angiogenesis Inhibitor	6	6
Chemotherapy+Immunotherapy	40	41
Immunotherapy	27	27
Creatinemia (µmol/l)		
Mean (sd)	65.5 (19.6)	64.6 (15.15)
Median	62.3	64.3
min-max	43.9-92.8	45.6-83.7

DP: Differential Privacy; BW: Before the wave; AW: After the wave; BMI: Body Mass Index; sd: Standard Deviation; ε: privacy budget

Discussion

In this study, a federated analysis with and without differential privacy was performed on a real-world study in oncology. The real-world study showed consistent results in both settings. In particular, identical conclusions were drawn regarding the lack of difference in first-line treatment duration and disease progression at 24 months in non-small cell lung cancer patients treated before or after the first wave of COVID-19.

Real world deployment in 3 centers has shed light on some key aspects in terms of security. Regarding tools, DataSHIELD offers a lot of flexibility as it allows the use of custom libraries such as dsPrivacy implemented as part of this study. The downside is that community libraries are not always well maintained and bugs can be encountered (e.g. dsStats had to be fixed manually). DataSHIELD builds upon Opal, a convenient data management application which comes with straightforward but very satisfactory user management and makes it easy to configure access to specific data to specific users. Opal is also secured by design and enforces many good security practices like CSRF (Cross-Site Request Forgery) and HTTPS usage. However, available docker images to deploy it have several major known security vulnerabilities and don’t seem regularly updated. Regarding network configuration, setting up strong security standards across all sites has

proved challenging. An IpSec bridge was set up to enable secure communication between the central node and the hospitals. Communication between all stakeholders was a key factor for this step and formal processes (network schema, requirements formalization, debugging methods, etc) proved decisive in order to facilitate discussions. Regarding certificates to secure communications inside the private network, the primary intention was to use certificates signed by hospitals internal certificate authority (CA) because it wouldn't match our security requirements so certificates signed by public CA have been claimed. Some hospitals provided a Sertigo certificate but it was not natively recognized by the central node server. Another hospital used instead a public certificate generated by the central node manager (Arkhn). This last solution has proven to be the best solution in terms of efficiency and security, even if the domain names covered by these certificates are not representing the actual ownership of hosts.

Regarding data standards, we chose FHIR since it appeared to be the most mature health standard in oncology at the beginning of this project [15, 16]. However, OMOP is developing extensions for oncology and is rapidly closing the gap [18, 19]. As the FHIR standard corresponds to json files that are not very suitable for analytical purposes, we anticipate that moving to OMOP will be more convenient for future analyses.

Overall, our experience suggests that the most time and work intensive parts are building harmonized data models locally and setting up the network infrastructure. As both these steps are agnostic of the study considered, we anticipate that they can be easily leveraged to scale the number of studies done on this infrastructure.

On the differential privacy side, it has been decided not to perform linear regression and Cox models with DP because implementing it with a moderate privacy budget was deemed too complex. Privacy budgets used in this study (especially for computing the *P*-values) are quite high compared to the literature [20, 14], in order to achieve reasonable accuracy on differentially private results but we have shown that we are able to derive similar results compared to plain federated analytics. As underlined above, the appropriate level of privacy is highly dependent on the context considered.

Here, given that this study was already compliant to the French MR-004 methodology without DP, the use of DP was done by prioritizing utility over privacy. In other contexts, especially if DP becomes recognized as a legitimate and safe technology to process personal data, and hence benefits from a dedicated regulatory status, lower privacy budgets will probably be required (alongside privacy-focused pen tests). We have identified three directions to achieve this:

- Increase the number of patients per center. This is the most straightforward option to improve the privacy / accuracy trade-off, since the amount of noise to add to reach a certain privacy budget ε directly depends on the number of individuals. We have included close to 50 patients per center which is quite low especially for some budget intensive operations for which we estimate that a thousand patients would be adapted.

- Use central differential privacy. If the noise is added on the central node, it is computed considering the sum of the patients across all the centers and enables lower noise for the same privacy budget. This is especially powerful if many centers are participating in the federated analysis. However, this means that by default the central node can see the contribution of each center without any noise which is a privacy breach if it can't be trusted. A common solution is to implement secure aggregation [21], meaning that all contributions are hidden and are only disclosed after the aggregation operation, using cryptographic mechanisms such as secure multi-party computation [22], homomorphic encryption [23] or functional encryption [24]. Secure aggregation is not yet implemented in dsPrivacy and is left as future work.

- Improve the differential privacy mechanism. We have used pure or ε-differential privacy, which is a simple mechanism and which makes composition very simple, since the privacy budget of a sequence of operations is the sum of the budget of each operation. More complex mechanisms such as (ε, δ)-differential privacy [6] or Rényi differential privacy [25] provide more optimized composition properties [26] that allow for a tighter privacy budget management.

Conclusion

The DARAH project illustrates federated analytics (FA) are a practical method to conduct scientific projects while improving data privacy, by keeping patient data stored in the hospitals and leveraging their already existing data architecture. It highlights some key challenges to be anticipated and possible answers to ensure the success of this type of projects. It also shows that differential privacy (DP) can be used in addition to FA to improve privacy guarantees, but more experimentation is needed to develop guidelines and best practices, especially around the trade-off between accuracy and privacy. Finally, in an emerging ecosystem where tools for FA and DP are not yet well integrated, the dsPrivacy library will prove useful for researchers who want to explore privacy-friendly analysis methods.

Funding

This work was supported by private funding from Roche. Apart from this, the authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Ethics approval

This is an observational study. The research ethics committees of all three centers in Toulouse, Reims and Foch have confirmed that no ethical approval is required and have authorized this project on their patient data.

Patient information

All patients enrolled in this study were informed individually, and those who exercised their right to opt out were removed from the study.

Data Availability

All generic source code has been made available as open source in the dsPrivacy and is available at https://github.com/arkhn/dsPrivacy.
Specific code to reproduce the study will be made available upon reasonable request to the corresponding author.
The data underlying this article cannot be made publicly available due to the privacy of the patients involved in this study.

References

[1] Rieke N, Hancox J, Li W, Milletarì F, et al. The future of digital health with federated learning. *NPJ Digit Med* 2020 Sep;3:119.

[2] Gaye A, Marcon Y, Isaeva J, et al. DataSHIELD: taking the analysis to the data, not the data to the analysis. *Int J Epidemiol* 2014;43(6):1929–44.

[3] Jaddoe VWV, Felix JF, Andersen AMN, et al. The LifeCycle Project-EU Child Cohort Network: a federated analysis infrastructure and harmonized data of more than 250,000 children and parents. *Eur J Epidemiol* 2020;35(7):709–24.

[4] Prokosch HU, Acker T, Bernarding J, et al. MIRACUM: Medical Informatics in Research and Care in University Medicine. *Methods Inf Med* 2018;57(S 01):e82–91.

[5] Gruendner J, Wolf N, Tögel L, et al. Integrating Genomics and Clinical Data for Statistical Analysis by Using GEnome MINIng (GEMINI) and Fast Healthcare Interoperability Resources (FHIR): System Design and Implementation. *J Med Internet Res* 2020;22(10):e19879.

[6] Shokri R, Stronati M, Song C,et al. Membership Inference Attacks Against Machine Learning Models. *2017 IEEE Symposium on Security and Privacy* (SP) 2017:3–18.

[7] Fredrikson M, Jha S, Ristenpart T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security 2015:1322–33.

[8] Dwork, Cynthia, et al. Our data, ourselves: Privacy via distributed noise generation. Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques 2006.

https://link.springer.com/chapter/10.1007/11761679_29 (accessed 29 Nov 2023)

[9] Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy. *FNT in Theoretical Computer Science* 2013;9(3–4):211–407

[10] Accessed online: https://github.com/google/differential-privacy (accessed 29 Nov 2023)

[11] Accessed online: https://github.com/OpenMined/PyDP (accessed 29 Nov 2023)

[12] Accessed online: https://github.com/opendp/opendp (accessed 29 Nov 2023)

[13] Accessed online: https://github.com/brubinstein/diffpriv (accessed 29 Nov 2023)

[14] Hsu J, Gaboardi M, Haeberlen A, Khanna S, et al. Differential Privacy: An Economic Method for Choosing Epsilon. 2014 IEEE 27th Computer Security Foundations Symposium 2014:398–410

[15] Accessed online: https://build.fhir.org/ig/InstitutNationalduCancer/ImplementationGuide_OsirisFHIR/(accessed 29 Nov 2023)

[16] Accessed online: https://build.fhir.org/ig/HL7/fhir-mCODE-ig/ (accessed 29 Nov 2023)

[17] Accessed online:

https://search.r-project.org/CRAN/refmans/DPpack/html/tableDP.html (accessed 29 Nov 2023)

[18] Belenkaya R, Gurley MJ, Golozar A, et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. *JCO Clin Cancer Inform* 2021;5:12–20.

[19] Accessed online: https://github.com/OHDSI/OncologyWG/wiki (accessed 29 Nov 2023)

[20] Lee J, Clifton C. How Much Is Enough? Choosing ε for Differential Privacy. Information Security: 14th International Conference, ISC 2011:325–40.

[21] Bonawitz K, Ivanov V, Kreuter B, et al. Practical Secure Aggregation for Federated Learning on User-Held Data. arXiv 2016;http://arxiv.org/abs/1611.04482 (accessed 29 Nov 2023)

[22] Donald Beaver. Efficient multiparty protocols using circuit randomization. Advances in Cryptology — CRYPTO ’91 199:420–432.

[23] Paillier P. Public-Key Cryptosystems Based on Composite Degree Residuosity Classes. Advances in Cryptology — EUROCRYPT ’99 1999:223–38

[24] Boneh D, Sahai A, Waters B. Functional Encryption: Definitions and Challenges. Theory of Cryptography; 2011:253–73

[25] Mironov I. Rényi Differential Privacy. IEEE 30th Computer Security Foundations Symposium (CSF) 2017:263–75.

[26] Kairouz P, Oh S, Viswanath P. The Composition Theorem for Differential Privacy . arXiv 2015.https://arxiv.org/abs/1311.0776 (accessed 29 Nov 2023)

[27] Hripcsak G, Duke JD, Shah NH,et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015;216:574–8.

[28] Alvim MS, Andrés ME, Chatzikokolakis K,et al. Differential Privacy: On the Trade-Off between Utility and Information Leakage. Formal Aspects of Security and Trust 2012:39–54

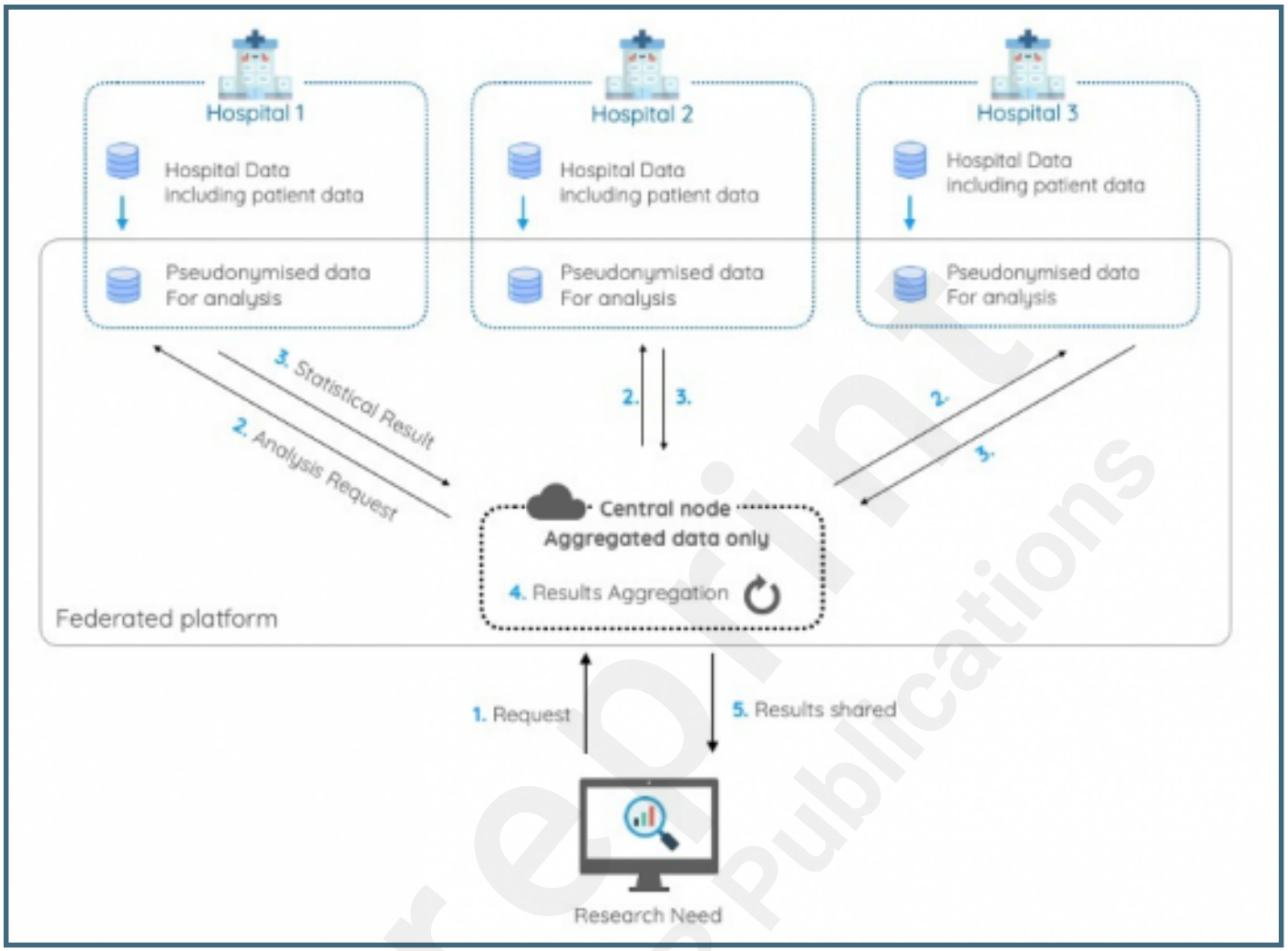
Abbreviations

FA: Federated analytics
DP: Differential Privacy
DARAH: Distributed Analytics for Research in Hospitals
NSCLC: Non-Small Cell Lung Cancer
BW: before the first wave
AW: after the first wave

Supplementary Files

Figures

Federated architecture diagram.



Flowchart.