Does federated analytics preserve statistical and scientific value of real-world data?

David Pau (1), Ismael Diabate (2), Lukasz Kaczmarek (3), Jacek Chmiel (4), Romain Jegou (5), Mathieu Boucher (5), Charles Monteil (2), Camille Bachot (1)

(1) Medical Data Center, Boulogne-Billancourt, France (2) PIIX, Boulogne-Billancourt, France (3) PIIX, Basel, Switzerland (4) Avenga, Poland (5) Keyrus Life Science, Nantes, France

Introduction

It is difficult (and sometimes impossible) to pull out data from different hospitals, and always a long journey to process necessary data privacy authorizations. The sum of all these data is a very valuable asset, which may benefit all partners involved (1) (2). Federated Analytics (FA) allows to generate aggregated results from separate data sources without data transfer and is respectful of data privacy and property of each data source.

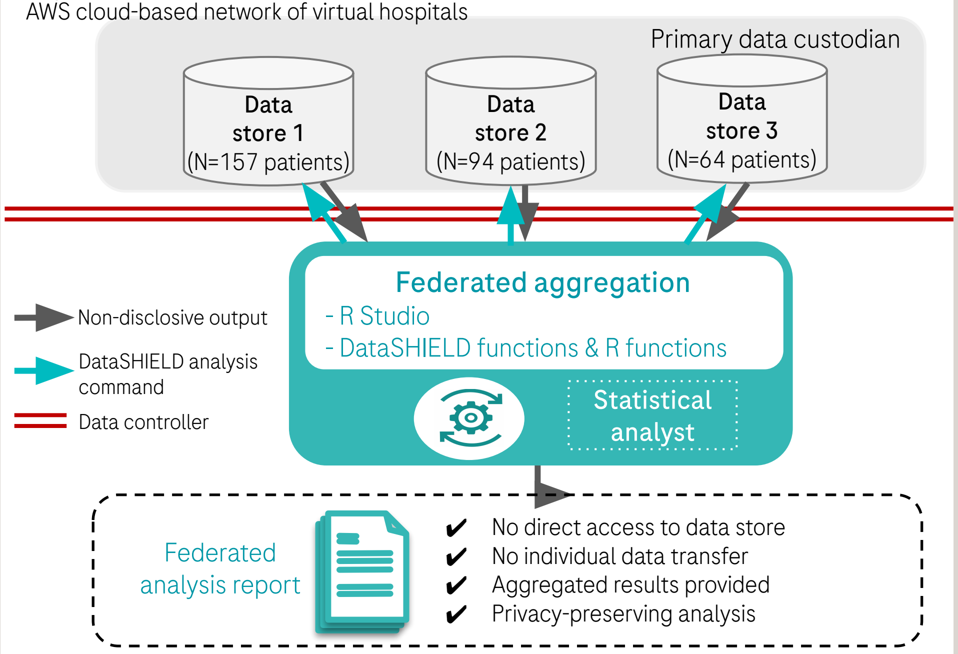
The area of FA is new and developing very rapidly, there is a lot of research, and more and more are ready to use patterns and tools. Internal benchmark on FA was performed in July 2021. The goal of this research was to create a map of capabilities of existing FA and data processing technologies available as open-source products. DataSHIELD open-source product has been selected for this project (3).
DataSHIELD is an infrastructure and series of R packages that enables the remote and non-disclosive analysis of sensitive research data (Figure 1).

Project objective

The goal is to evaluate the DataSHIELD tool on real-world data, using anonymized data of a non-interventional study. The primary objective is to understand which types of statistical analysis are possible to perform and which are not yet possible. Common statistical analyses produced on the raw data will be reproduced and compared using FA DataSHIELD environment.

Figure 1 : Federated analytics environment

AWS cloud-based network of virtual hospitals

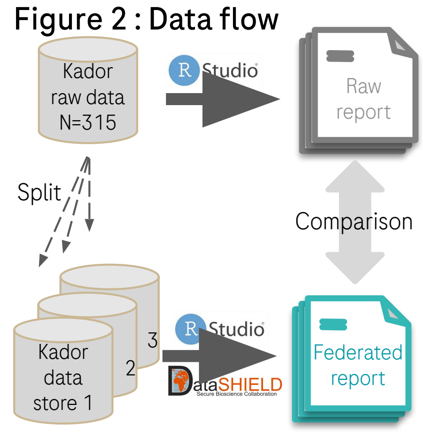

Primary data custodian
Data store 1 (N=157 patients)
Data store 2 (N=94 patients)
Data store 3 (N=64 patients)
Federated aggregation
- R Studio
- DataSHIELD functions & R functions
Statistical analyst
Non-disclosive output
DataSHIELD analysis command
Data controller
Federated analysis report

- ✔ No direct access to data store
- ✔ No individual data transfer
- ✔ Aggregated results provided
- ✔ Privacy-preserving analysis

Project method

Anonymised data of the KADOR longitudinal real-world study have been selected for this project. The selected KADOR study datasets has been randomly splitted into 3 databases of different size (N1=157 patients, N2=94 and N3=64), stored in a data platform which integrates DataSHIELD. Common statistical analysis will be reproduced (114 items) using DataSHIELD and R statistical functions and compare to the report obtained from the raw data (Figure 2).

Figure 2 : Data flow



Data transformation:
Analysis variable derivation (class, duration, criteria), data selection
Univariate and bivariate descriptive statistics:
continuous parameters (mean, median, Q1/Q3 and min/max) and dichotomous parameters (each modality with number/% of patients).
Survival analysis:
Kaplan-Meier, Univariate/Multivariate Cox model
Logistic analysis:
Univariate/Multivariate model
Correlation matrix:
Chi-squared test, ANOVA

Table 1 : Descriptive statistics comparison

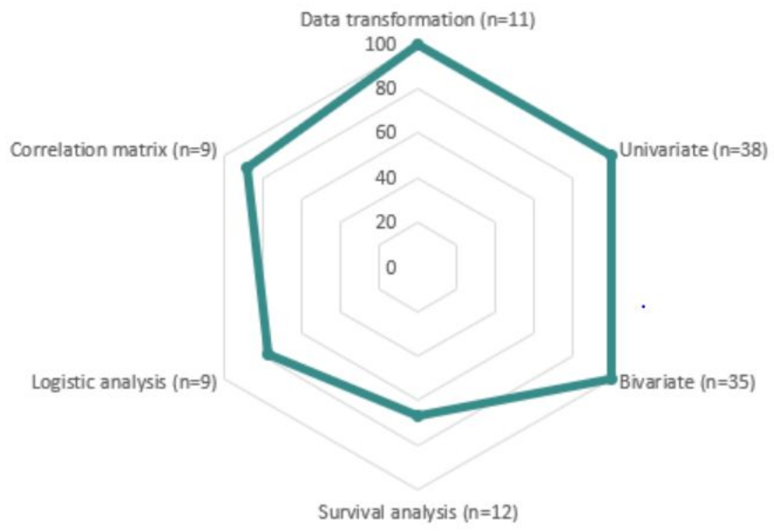
		Raw (N=315)	Federated (N=315)
Age at treatment initiation (years) *	N	303	303
	Mean (SD)	52.23 (11.82)	52.23 (11.82)
	Median (Q1;Q3)	52.0 (43.0; 60.5)	52.1 (43.4; 60.8)
	Min - Max	29 - 77	NA - NA
	Missing	12	12
SBR grade **	N	304	304
	SBR I	5 (1.6%)	5 (1.6%) ***
	SBR II	139 (45.7%)	139 (45.7%)
	SBR III	154 (50.7%)	154 (50.7%)
	Ungradable	6 (2.0%)	6 (2.0%) ***
	Missing	11	11
pCR results **	N	315	315
	pCR	127 (40.3%)	127 (40.3%)
	No pCR	188 (59.7%)	188 (59.7%)
	Missing	0	0

Differences for quantiles due to aggregation approximation

Min/max values not provided for federated analysis (disclosive)

* DataSHIELD functions use for continuous summary: dh.getStats, ds.meanSdGp.
** DataSHIELD functions use for categorical summary: dh.getStats, ds.table.
*** With DataSHIELD, the minimum non-zero cell count allowed in any cell if a contingency table is per default set to 3 to prevent disclosive results. Data triplication has been performed as a work around in DataSHIELD server side in order to retrieve modality with < 3 observations in at least one data store.

Figure 3 : Current percentage of statistical analyses from the raw report successfully reproduced using DataSHIELD by type of analysis (n=114 items)



Conclusion

This project aims to evaluate if scientific value and statistical results are maintained with FA programming analysis environment.
DataSHIELD open-source product enables usual aggregated analyses from separated data sources without individual data transfer and by preserving data privacy and property. Analyses can be performed using built-in DataSHIELD functions (e.g. dsBase, dsHelper) and/or freely available community packages (e.g. dsSurvival).
Most of the planned common statistical analysis have been reproduced with similar results (Figure 3) using available built-in DataSHIELD functions (data transformation, univariate/bivariate descriptive analyses, correlation matrix, univariate logistic analysis).
However, some analysis (survival analysis) are not easily reproducible (especially Kaplan-Meier curves) using built-in functions and need further GDPR compliance evaluation before specific functions to be developed.

References:

(1) Sheller MJ, Edwards B, Reina GA, Martin J, Pati S, Kotrotsou A, Milchenko M, Xu W, Marcus D, Colen RR, Bakas S. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep. 2020 Jul 28;10(1):12598. doi: 10.1038/s41598-020-69250-1. PMID: 32724046; PMCID: PMC7387485.

(2) Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated Learning for Healthcare Informatics. J Healthc Inform Res. 2021;5(1):1-19. doi: 10.1007/s41666-020-00082-4. Epub 2020 Nov 12. PMID: 33204939; PMCID: PMC7659898.

(3) DataSHIELD: taking the analysis to the data, not the data to the analysis. International Journal of Epidemiology (2014).

Disclosure: this work is funded by F. Hoffmann-La Roche Ltd