ISCB, Milan, 2023

Federated Analysis of Multiple Data Sources – Centralised Analysis of Decentralised Databases Using the DataSHIELD Software

Wilmar Igl[1] & Stefano Viaggi[2]

[1] ICON PLC, Sweden; [2] ICON PLC, Italy

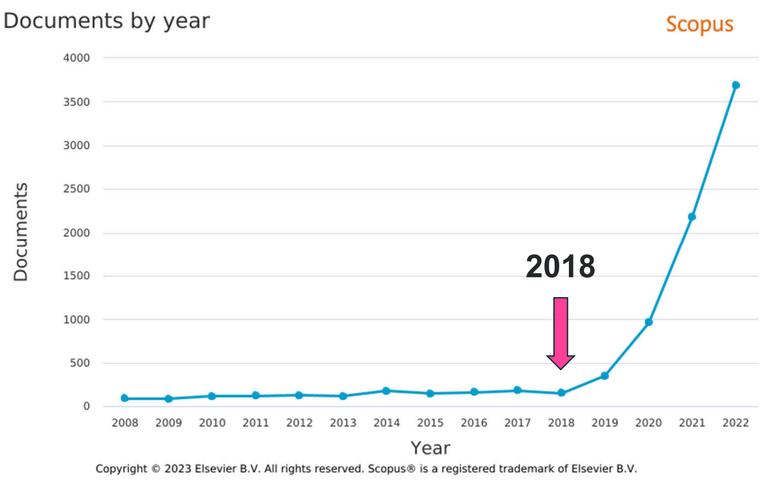
Introduction

Federated Analysis (FA) has received rapidly increasing attention over the last five years in the scientific literature (**Figure 1**).

FA describes the centralized analysis of decentralized databases while preserving the privacy of personal data [1, 2].

FA was proposed for the identification of rare adverse events in international post-marketing studies to overcome legal barriers [3].

Figure 1. Publications on Federated Analysis over Time

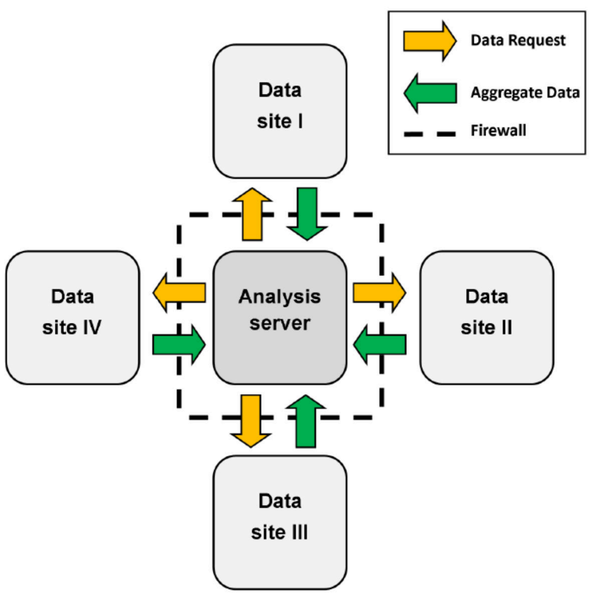


Key Concepts

Federated Databases

- centralized analysis
- decentralized databases
- standardized Advanced Programming Interface (API)
- site-specific data control ("Firewall"), e.g. no access to individual subject data

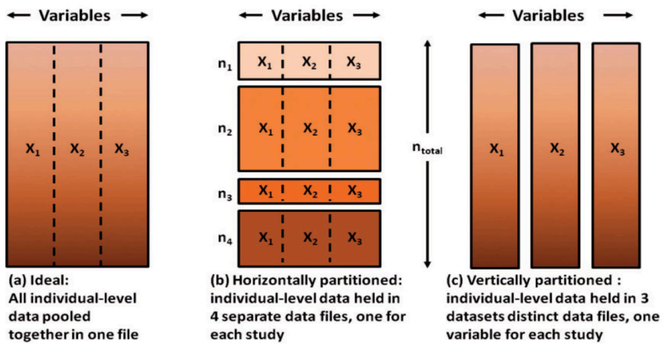
Figure 2. The Concept of Federated Analysis



Data Structures

1. Ideal:
   All patients and all variables in one database
2. Horizontally partitioned:
   All variables in all databases, but patients in different databases
3. Vertically partitioned:
   All patients in all databases, but variables in different databases

Table 1. Ideal, horizontally partitioned, & vertically partitioned data structures



(a) Ideal: All individual-level data pooled together in one file

(b) Horizontally partitioned: individual-level data held in 4 separate data files, one for each study

(c) Vertically partitioned : individual-level data held in 3 datasets distinct data files, one variable for each study

- Federated Analysis applicable to horizontally and vertically partitioned data, but current developments focus on horizontally partitioned data.

Statistical Concepts

Comparison with Conventional Analysis

Table 1. Comparison of Standard Analysis, Meta-Analysis, and Federated Analysis

	Standard Analysis	Meta-Analysis	Federated Analysis
Architecture	centralized analysis, centralized data	(de)centralized analysis, decentralized data	centralized analysis, decentralized data
Data	id-level	group-level	id|group-level
Statistics	full	limited (fixed vs random)	GLM, Cox PH, … (or in development)
Computation	id ↔ model	id → group ↔ model	id → group ↔ model id ← model
Privacy	low	high	high

Statistical Principle: Decomposition of Statistical Loss

- A **global loss function** is decomposed into the sum of the weighted combination of multiple local loss functions. [1]
- **Statistical loss** is a measure of the costs of the statistical errors in the estimation of a parameter used to estimate its optimal value (cf likelihood function).

Equation 1. Statistical Loss

$$\min_{\phi} \mathcal{L}(X;\phi) \quad \text{with} \quad \mathcal{L}(X;\phi) = \sum_{k=1}^{K} w_k \, \mathcal{L}_k(X_k;\phi)$$

- ϕ : parameter
- X : unavailable complete data
- $\sum_{k=1}^{K} w_k \, \mathcal{L}_k(X_k;\phi)$: sum of local loss functions L_k with weight w_k

Example: Generalized Linear Models

- **Linear Predictor:** $\eta_i := g(\mu_i) = \boldsymbol{\beta}^T \boldsymbol{x}_i$
- **Iterative Reweighted Least Square Algorithm**

$$\boldsymbol{\beta}_{t+1} = \boldsymbol{\beta}_t + \boldsymbol{I}(\boldsymbol{\beta}_t)^{-1} \boldsymbol{s}(\boldsymbol{\beta}_t)$$

- **Information Matrix:**

$$\boldsymbol{I}(\boldsymbol{\beta}_t) = \boldsymbol{X}^T \boldsymbol{W}_t \boldsymbol{X}$$

$$\boldsymbol{I}(\boldsymbol{\beta}_t) = \sum_{i=1}^{N} w_{ii}(t) \, \boldsymbol{x}_i \boldsymbol{x}_i^T$$

- **Score Function**

$$\boldsymbol{s}(\boldsymbol{\beta}_t) = \boldsymbol{X}^T \boldsymbol{W}_t (\boldsymbol{Y} - \boldsymbol{\mu}(t)) g'(\boldsymbol{\mu}(t))$$

$$\boldsymbol{s}(\boldsymbol{\beta}_t) = \sum_{i=1}^{N} (y_i - \mu_i(t)) g'(\mu_i(t)) w_{ii}(t) \boldsymbol{x}_i$$

- **Convergence**:

$$\frac{|D_r - D_{r-1}|}{D_r + 0.1} < 10^{-8}$$

Available Statistical Functionality (Present)

- Descriptive statistics and visualizations
- Inference statistics [3, 5, 6]
  - (Meta-analysis)
  - Generalized Linear Models
  - Cox Proportional Hazards Model

DataSHIELD software

- DataSHIELD software: Data aggregation through anonymous Summary-statistics from Harmonized Individual levEL Databases (DataSHIELD) [5, 6] (https://datashield.org)
- Multi-component software stack, e.g. OPAL, ROCK/R, Mango/MySQL
- Official R packages (https://cran.datashield.org)
  - Client Packages: dsBaseClient
  - Server Packages: dsBase, opaladmin
  - Testing (serverless implementation): DSLite
- Multiple community packages, e.g.:
  dsOmics, dsExposome, dsHelper, dsSurvival, dsMediation, dsSwissKnife, dsML, dsGeo, dsDanger, dsMicrobiome, dsQueryLibrary, dsBoltzmannMachines, dsMTL, dsSynthetic, dsClusterAnalysis

DataSHIELD software (cont.)

DataSHIELD Software Stack

- **Analyst (Client):** R/DataSHIELD packages (DSI, DSOpal, dsBaseClient)
- **Data Owner (Server):**
  - Data Warehouse: OPAL & Database: MANGO/MySQL
  - JAVA
  - R Server ROCK and Statistical Analysis System R
  - Webserver: NGINX (with TLS certificate)

DOCKER Technology

- OS-level virtualization to deliver software in packages called containers
- **Installation of the DataSHIELD software stack on a bare-bone Linux server in about 30 minutes [2]:**

```
SHELL> sudo docker-compose -f dsconf.yml up -d
```

Example: Post-Vac Syndrome

- **"Post-Vac Syndrome"**: Definition (here) as "Postviral fatigue syndrome/Myalgic encephalomyelitis" (PFS/ME) (ICD-10: G93.3, MedDRA 25.1: 10008874) after vaccination against COVID-19, also known as Myalgic encephalomyelitis/Chronic Fatigue Syndrome (ME/CFS)
- **Objective**: Descriptive analysis of ICD-10 G93.3 disease after vaccination with COVID-19 vaccines:
  Comirnaty (BioNTech and Pfizer), COVID-19 Vaccine (Valneva), Nuvaxovid (Novavax), Spikevax (Moderna), Vaxzevria (AstraZeneca), Jcovden (Janssen), VidPrevtyn Beta (Sanofi Pasteur), Bimervax (HIPRA Human Health SLU)
- **Data**: VigiBase database (WHO, March 2023 [7]) with data from Europe and America (2021/2022), which was split and stored into separate databases to allow Federated Analysis
- **Statistical Analysis**: Absolute Frequencies (AF, "counts") and Relative Frequencies (RF, "reporting rate") of Individual Case Safety Report (ICSR) in the Federated Database (FA) based on data from Europe (EU) and America (AM)
- **Results:**
  - EU: AF(G93.3) = 1074, AF = 2,256,738, RF = 0.48‰
  - AM: AF(G93.3) = 679, AF = 1,676,488, RF = 0.41‰
  - FA: AF(G93.3) = 1753, AF = 3,933,226, RF = 0.45‰
- **Conclusion**:
  - FA can be successfully applied without sharing confidential patient data
  - No causal conclusion of Covid-19 vaccination on PFS/ME as adverse event.

Conclusions

- **Federated Analysis** offers centralized analysis using the (full) information from de-centralized individual data with high-levels of privacy
- **Important statistical models** were re-formulated and are available as software
- **DataSHIELD software** is easily deployed using Docker technology

References

1. Rieke, N. et al. (2020). The future of digital health with federated learning. npj Digital Medicine, 3(1), 1. https://doi.org/10.1038/s41746-020-00323-1
2. Wilson, R.C., et al. (2017). DataSHIELD – New Directions and Dimensions. Data Science Journal, 16, 21
3. Gedeborg, R., Igl, W., Svennblad, B., Wilén, P., Delcoigne, B., Michaëlsson, K., Ljung, R., & Feltelius, N. (2022). Federated analyses of multiple data sources in drug safety studies. Pharmacoepidemiology and Drug Safety. https://doi.org/10.1002/pds.5587, Supporting Information (Original Technical Reports, .zip): https://tinyurl.com/bwb75efu
4. Igl, W. (2023). Federated Analysis with R/DataSHIELD – Installation, Data Import, Data Analysis [Tutorial]. https://wilmarigl.de/?p=424
5. Jones, E., et al. (2012). DataSHIELD – shared individual-level analysis without sharing data: a biostatistical perspective. Norwegian Journal of Epidemiology 21(2), 231-239.
6. Michael Wolfson et al. (2010). DataSHIELD: resolving a conflict in contemporary bioscience - performing a pooled analysis of individual-level data without sharing the data. International Journal of Epidemiology, 39(5), 1372–1382. Supplementum (with detailed statistical examples)
7. Lindquist M. (2008). VigiBase, the WHO Global ICSR Database System: Basic Facts. Drug Information Journal. 42(5):409-19. https://who-umc.org
   Note: VigiBase is the WHO global database of reported potential side effects of medicinal products, developed and maintained by Uppsala Monitoring Centre. The information comes from a variety of sources, and the probability that the suspected adverse effect is drug-related is not the same in all cases. The information does not represent the opinion of the UMC or the World Health Organization.