Evaluation of statistical software for federated analysis of multi-site real world studies

Fiona Grimson (fiona.grimson@iqvia.com)[1], Nicolas Niklas[2], Ruben Hermans[1], Lorena Cirneanu[1], Benedikt Maissenhaelter[2], Joseph Kim[1,3]

[1]NEMEA Centre of Excellence for Retrospective Studies, Real-World Insights, IQVIA, London, UK; [2]IQVIA, Frankfurt, Germany; [3]Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine

IMS Health & Quintiles are now IQVIA™

Key Takeaways

Federated analysis (FA) is an alternative to pooling individual patient data (IPD) or to meta-analysis that:

- Gives analytical results that are **equivalent to pooled IPD,** and
- **Fully preserves data privacy,** as IPD is never shared outside each contributing site

While there is potential for FA to be effective for real world evidence studies, and there have been many developments in the federation of analytical methods in recent years, there are still gaps which make FA difficult to execute in practice:

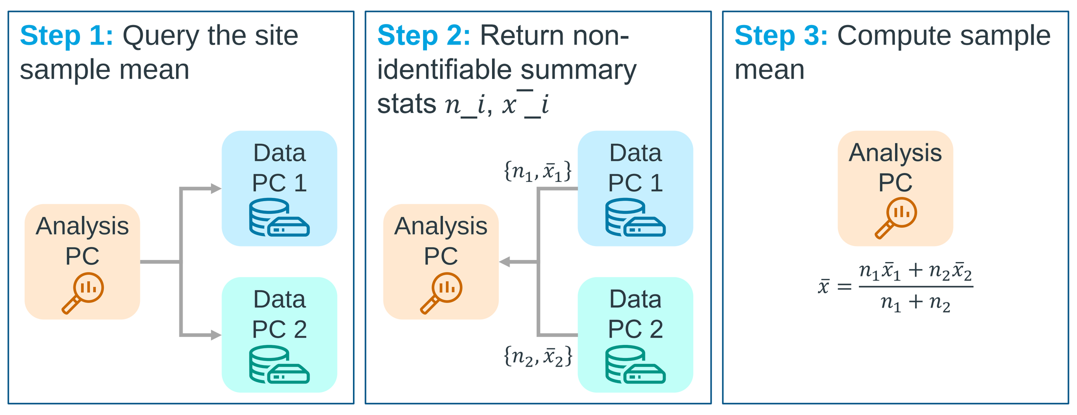
- **Only a limited number of analytical methods have been federated**
- **Fragmented landscape of federated software solutions**

Substantial software development is required before a typical real world study can be readily performed with FA.

How are federated analyses for some common statistical methods executed?

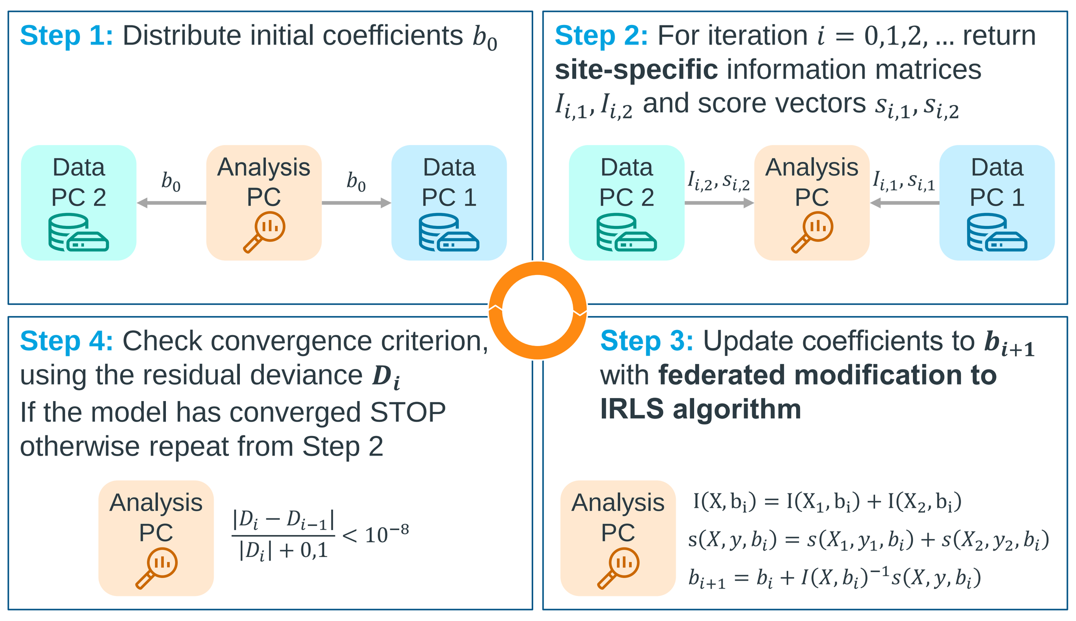
Descriptive statistics: For instance, the sample mean. The sufficient statistics for the sample mean $\bar{x}$ across s sites are $\{n_i, \bar{x}_i\}$ for $i = 1, \ldots, s$, because $\bar{x} = \frac{\sum_{i=1}^{s} n_i \bar{x}_i}{\sum_{i=1}^{s} n_i}$. The site contributions to the sample mean do not retain any identifiable patient data, and the result is exactly equivalent to pooled data.

A simple federated analysis – sample mean



Generalised linear models: This is made possible by a modification of the iterated reweighted least-squares algorithm which iteratively requests non-identifying summary statistics from each site. An example of this is logistic regression. There can be some numerical differences in the model coefficients, as expected from a numerical optimization procedure.[2,3]

A more complex federated analysis – logistic regression



Cox Proportional Hazards models: The federated analysis results are equivalent to pooled analysis under Breslow's partial likelihood assumption.[4]

Patient matching: Patient similarity is measured by assigning context-specific binary hash codes to patients, and computing the hamming distance counting the number of positions where the hash codes differ. The creation of the patient feature vector to be hashed is selected by the researcher, and would require similar consideration to feature selection in other matching contexts such as propensity score matching.[5]

Federated Analysis (FA) for multi-site real world database studies

In real world evidence (RWE), multi-site studies are needed to obtain sufficiently large and representative patient cohorts from electronic medical records databases. Typical approaches to multi-site analyses are either to

- **Pool individual patient data** (IPD) into a single research database (Figure 1), or
- **Perform a meta-analysis** using site-level summary statistics (Figure 2)

Both methods pose challenges to researchers:

- Data pooling requires contributing sites to give their IPD to a third party, which poses both regulatory and trust challenges due to data privacy laws including EU GDPR
- Meta-analysis does not make use of IPD. Using IPD in a multi-site study increases the precision of analyses compared to meta-analysis, and can still incorporate appropriate weighting between contributing sites or subgroups as required

Federated analysis (Figure 3) is an alternative method:

- Harmonized datasets containing sensitive patient-level data are hosted securely at contributing sites, so sites retain complete control of their data
- There is full preservation of data privacy, as identifiable data is never shared outside the site[1]
- Analyses are performed simultaneously across sites, and give results equivalent to those obtained with pooled IPD

FA is an effective when:

- Sites cannot, or do not wish to, share IPD
- The statistical methods required for the analysis can be federated
- The technology and software for a federated analysis is available

Figure 1: Pooled analysis

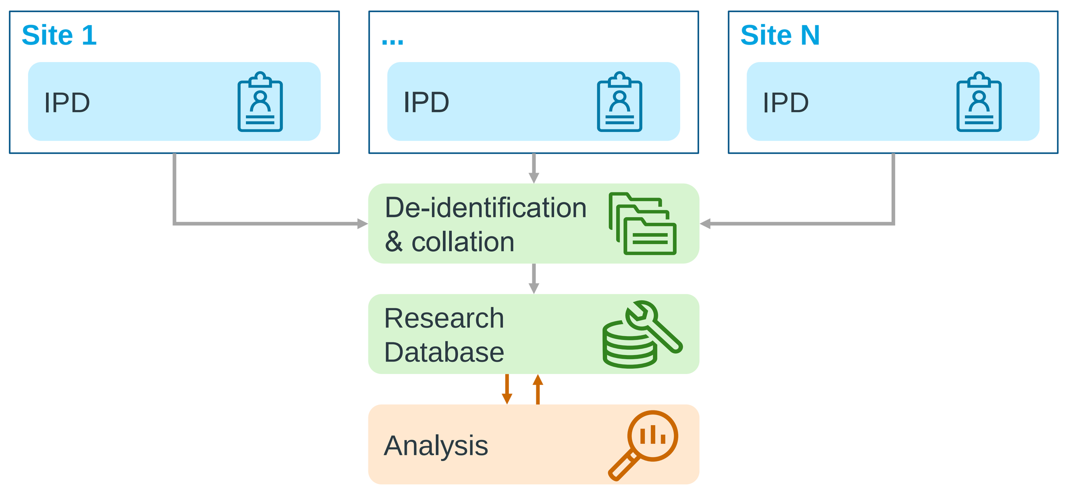


Figure 2: Meta-analysis

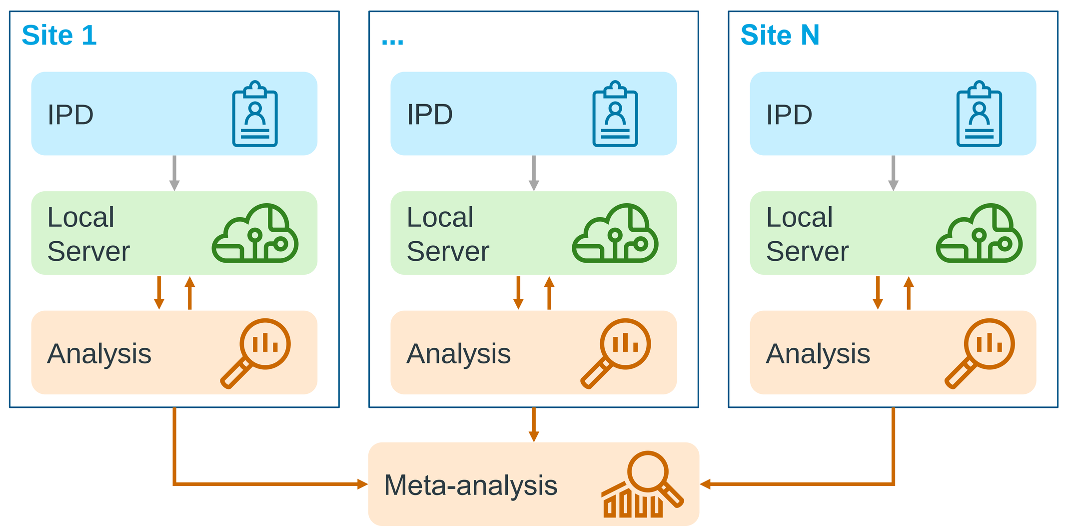
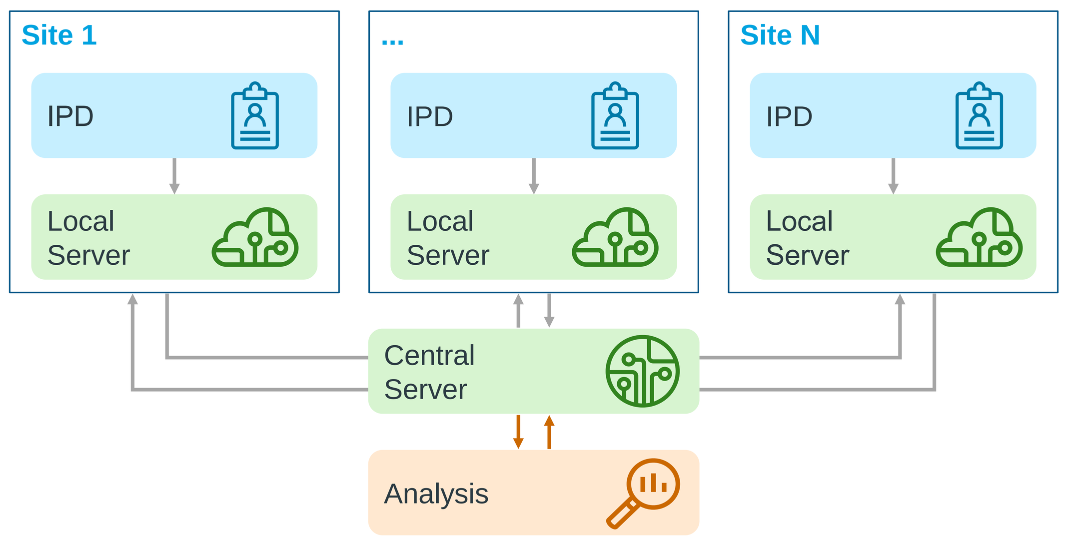


Figure 3: Federated Analysis



Review of FA Software for common statistical methods

Commonly used statistical methods for real world studies were identified through literature review of recent publications and consultation of senior real world data researchers.

For the analytical methods, a targeted literature review was conducted to identify:

1. Demonstration of federated execution;
2. Proof of federated results being exact and non-disclosive;
3. Availability of software either commercial or open source.

The review did not cover compliance with hospital information governance, RBAC, functionality of software, or other aspects.

The majority of the analytical methods identified have been federated, and many have software implementations, listed in Table 1.

Table 1: Software for common statistical methods

Method	Description	Proof that federated execution is possible	Software Implementation
Descriptive Statistics	Mean, Standard Deviation, IQR, Count, Percentages, Median, Min, Max, Contingency Tables, Proportion, Odds ratio	[7]	DataSHIELD [7]; NB: Statistics such as min, max, median may be disclosive
Visualization	Histogram, contour plot, heat map, scatter plot, box pot	[25]	DataSHIELD [7] NB: Plots such as scatter plots and box plots may be disclosive
Hypothesis Tests	T-test, Wilcoxon Rank Sum, Wilcoxon Signed Rank, Chi-square test	[6]	DataSHIELD [7], ShareMIND [8], RMIND [9], DataMole [10], SMC [11]
	Z-test, McNemar's Test, ANOVA, Fisher's Exact Test	[12]	DataSHIELD [7], SMC [11]
	Kruskall-Wallis, Spearman Correlation, Pearson Correlation	[16]	SCS [17]
	Kendall's Tau Test, Kolgomorov-Smirnov	[10]	DataMole [10]
	Cochran-Armitage	[18]	SMC [11]
Variable Selection	LASSO	[12]	SMC [11]
	AIC, BIC, Likelihood Ratio Test	[19]	DataSHIELD [7], WebGLORE [20], WebDISCO [21]
Models	Linear Regression, Logistic Regression, Generalised Linear Model, Generalised Linear Mixed Model	[22]	DataSHIELD [7]
	Generalised Estimating Equations	[23]	SMC [11]
Survival Models	Kaplan Meier	[13]	tranSMART [14], SmartR [15]
	Cox Proportional Hazards Regression	[21]	WebDISCO [21]
Matching	Non-Exact Matching	[24]	SAFTINet [24]
	Propensity Score Matching	[5], [26]	HD-PS Algorithm [26]
Machine Learning	Bayesian Neural Networks, Gaussian Process Models	[27]	PVI Framework [27]

The current state of FA for real world studies

Proof of Federation for statistical analyses:

For statistical modelling and testing, the algorithms for model fitting must be federated, and proved to be equivalent to non-federated. While many common models identified have been federated (Table 1), some commonly used methods have not. This includes:

- Imputation (e.g. multiple imputation, or repeated measures models)
- Some survival models (e.g. parametric survival models, competing risks)
- Mixed and hierarchical models, particularly random intercept and random coefficient models.
- Bayesian analysis (a framework called PVI (Partitioned Variational Inference) [27] has been developed for federated training of Bayesian neural networks, but more work needs to be done)

For data visualization, the federation of the method for collecting plot data can be straightforward, but the limitation is whether individual data values can be distinguished in the plot. For instance, a scatter plot would not be suitable, but a density plot could be produced. More federated data visualizations would be valuable.

Software Availability:

No federated software solution has all methods implemented and combing platforms would cumbersome. Many solutions require their own tool or architecture and are not readily compatible.

Providers of federated analytical tools must also be able to prove that analytical methods and outputs do not violate any data sharing regulations and guarantee results do not contain patient-identifiable data.

References

1. Budin-Ljonse, I et al, DataSHIELD: an ethically robust solution to multiple-site individual-level data analysis. Public Heath Genomics 2015;18:87-96
2. Wolfson M et al. DataSHIELD: resolving a conflict in contemporary bioscience – performing a pooled analysis of individual-level data without sharing the data. Int J Epidemiol 2010;39:1372–1382
3. Jones, E.M, et al. DATASHIELD - shared individual-level analysis without sharing the data: a biostatistical perspective. Nordsk Epidemiologi 2012; 21 (2) 231-239
4. Lu CL et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. J Am Med Inform Assoc 2015; 22(6)1212-9
5. Lee J et al. Privacy-preserving patient similarity learning in a federated environment: development and analysis. JMIR Med Inform 2018;6(2):e20
6. Bogdanov D., Kamm L., Laur S., Pruulmann-Vengerfeldt P., Talviste R., Willemson J. (2014) Privacy-Preserving Statistical Data Analysis on Federated Databases. In: Preneel B., Ikonomou D. (eds) Privacy Technologies and Policy. APF 2014. Lecture Notes in Computer Science, vol 8450. Springer, Cham
7. DataSHIELD: taking the analysis to the data, not the data to the analysis, International Journal of Epidemiology, 2014
8. ShareMIND: Bogdanov, D., Laur, S., & Willemson, J. (2008). Sharemind: a framework for fast privacy-preserving computations. IACR Cryptology ePrint Archive.[4] RMIND
9. RMIND: Bogdanov, Dan & Kamm, Liina & Laur, Swen & Sokk, Ville. (2016). Rmind: a Tool for Cryptographically Secure Statistical Analysis. IEEE Transactions on Dependable and Secure Computing. PP. 1-1. 10.1109/TDSC.2016.2587623
10. DataMole: Mayo C, Conners S, Warren C, Miller R, Court L, Popple R. Demonstration of a software design and statistical analysis methodology with application to patient outcomes data sets. Med Phys. 2013 Nov;40(11) 111718. doi:10.1118/1.4824917. PMID: 24320426; PMCID: PMC3815052
11. SMC: https://sunfish-platform-documentation.readthedocs.io/en/latest/smc.html
12. Srinivasan S. (2017) Guide to Big Data Applications. Springer, Vol. 26
13. Herzinger, S., Grouès, V., Gu, W., Satagopam, V., Banda, P., Trefois, C., & Schneider, R. (2018). Fractalis: a scalable open-source service for platform-independent interactive visual analysis of biomedical data. GigaScience, 7(9), giy109. doi:10.1093/gigascience/giy109
14. Transmart: D. Athey, Brian. (2014). The tranSMART Open Data Sharing and Analytics Cloud Platform
15. SmartR: Herzinger, Sascha & Gu, Wei & Satagopam, Venkata & Eifes, Serge & Rege, Kavita & Barbosa Da Silva, Adriano & Schneider, Reinhard. (2017). SmartR: An open-source platform for interactive visual analytics for translational research data. Bioinformatics. 33. 10.1093/bioinformatics/btx137
16. Hoerbst A., Hackl W.O., de Keizer N. (2016). Exploring Complexity in Health: An Interdisciplinary Systems Approach: Proceedings of MIE2016. IOS Press, Vol. 228 of Studies in Health Technology and Informatics
17. SCS: Pawar P. S., Sajjad A., Dimitrakos T., Chadwick D. W. Security-as-a-Service in Multi-cloud and Federated Cloud Environments
18. Kamm L., Bogdanov D., Laur S., Vilo J. (2013). A new way to protect privacy in large-scale genome-wide association studies. Genetics and population analysis, Vol. 29, no. 7, 886-893
19. Her Q. L., Vilk Y., Young J., Zhang Z., Malenfant J., Malek S., Toh S. A distributed regression analysis application based on SAS software
20. Jiang, W., Li, P., Wang, S., Wu, Y., Xue, M., Ohno-Machado, L., & Jiang, X. (2013). WebGLORE: a web service for Grid LOgistic REgression. Bioinformatics (Oxford, England), 29(24), 3238–3240. doi:10.1093/bioinformatics/btt559
21. Lu, C. L., Wang, S., Ji, Z., Wu, Y., Xiong, L., Jiang, X., & Ohno-Machado, L. (2015). WebDISCO: a web service for distributed cox model learning without patient-level data sharing. Journal of the American Medical Informatics Association: JAMIA, 22(6), 1212–1219. doi:10.1093/jamia/ocv083
22. Wolfson, M., Wallace, S. E., Masca, N., Rowe, G., Sheehan, N. A., Ferretti, V., … Burton, P. R. (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. doi:10.1093/ije/dyq111
23. El Emam, K., Samet, S., Arbuckle, L., Tamblyn, R., Earle, C., & Kantarcioglu, M. (2012). A secure distributed logistic regression protocol for the detection of rare adverse drug events. Journal of the American Medical Informatics Association: JAMIA, 20(3), 453–461. doi:10.1136/amiajnl-2011-000735
24. Dewri R., Ong T., Thurimella R. (2016). Linking Health Records for Federated Query Processing. Proceedings on Privacy Enhancing Technologies, Vol. 3, 4-23.[23] El Emam, K., Samet, S., Arbuckle, L., Tamblyn, R., Earle, C., & Kantarcioglu, M. (2012). A secure distributed logistic regression protocol for the detection of rare adverse drug events. Journal of the American Medical Informatics Association: JAMIA, 20(3), 453–461. doi:10.1136/amiajnl-2011-000735
25. Wilson, R. C., Butters, O. W., Avraam, D., Baker, J., Tedds, J. A., Turner, A., … Burton, P. R. (2017). DataSHIELD – New Directions and Dimensions. Data Science Journal, 16, 21. DOI: http://doi.org/10.5334/dsj-2017-021
26. Rassen JA, Schneeweiss S. Using high-dimensional propensity scores to automate confounding control in a distributed medical product safety surveillance system. Pharmacoepidemiol Drug Saf. 2012;21: 41–49. 10.1002/pds.2328
27. Bui, T. D., Nguyen, C. V., Swaroop, S., & Turner, R. E. (2018). Partitioned Variational Inference: A unified framework encompassing federated and continual learning. arXiv preprint arXiv:1811.11206.