REVIEW

Federated analyses of multiple data sources in drug safety studies

Rolf Gedeborg | Wilmar Igl | Bodil Svennblad | Peter Wilén | Bénédicte Delcoigne | Karl Michaëlsson | Rickard Ljung | Nils Feltelius

Abstract

Purpose: Studies of rare side effects of new drugs with limited exposure may require pooling of multiple data sources. Federated Analyses (FA) allow real-time, interactive, centralized statistical processing of individual-level data from different data sets without transfer of sensitive personal data.

Methods: We review IT-architecture, legal considerations, and statistical methods in FA, based on a Swedish Medical Products Agency methodological development project.

Results: In a review of all post-authorisation safety studies assessed by the EMA during 2019, 74% (20/27 studies) reported issues with lack of precision in spite of mean study periods of 9.3 years. FA could potentially improve precision in such studies. Depending on the statistical model, the federated approach can generate identical results to a standard analysis. FA may be particularly attractive for repeated collaborative projects where data is regularly updated. There are also important limitations. Detailed agreements between involved parties are strongly recommended to anticipate potential issues and conflicts, document a shared understanding of the project, and fully comply with legal obligations regarding ethics and data protection. FA do not remove the data harmonisation step, which remains essential and often cumbersome. Reliable support for technical integration with the local server architecture and security solutions is required. Common statistical methods are available, but adaptations may be required.

Conclusions: Federated Analyses require competent and active involvement of all collaborating parties but have the potential to facilitate collaboration across institutional and national borders and improve the precision of postmarketing drug safety studies.

KEYWORDS
adverse drug reactions, decentralised analysis, federated analysis, pooled analysis, post-authorisation safety studies, product surveillance, postmarketing, registers

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2022 The Authors. Pharmacoepidemiology and Drug Safety published by John Wiley & Sons Ltd.
1 | INTRODUCTION

Characterising rare adverse drug reactions is a common and challenging regulatory concern, especially for new drugs and drugs with limited exposure. The need to combine multiple data sources for such analyses is becoming increasingly important, but combining different sources of individual-level data is a complex process, especially when different countries are involved.

The rapid development of the COVID-19 vaccines and need for a fast implementation of national vaccination programs has highlighted the need to generate timely post-approval safety data to detect potential uncommon adverse reactions. Having agreements and technical arrangements in place for such pooling of data from multiple sources and countries would greatly facilitate the generation of timely study results.

Federated analysis (FA) is a technique that may facilitate a centralised combined analysis of multiple decentralised data sources without requiring actual data merging. We review the findings in a FA development project conducted by the Swedish Medical Products Agency covering computer engineering aspects, limitations and opportunities of statistical methods, the legality and regularity of FA involving the processing of personal data, and tools for validating implementation.

The focus is on FA that can generate identical results as individual level pooling of data in comparative epidemiological studies requiring regression analysis to control for confounding.

2 | WHAT ARE FEDERATED ANALYSES?

Federated analysis allow real-time, interactive, central statistical analyses in a system of federated databases, without transferring individual level data outside the protective security mechanisms of the original host system (Figure 1). It is therefore an apparently attractive alternative to physical merging of data and there are several examples of initiatives implementing different forms of FA for pharmacoepidemiology. That the federated approach provides identical results compared to analyses of physically merged data has been demonstrated for commonly used statistical models, for example, Generalized Linear Regression models.

In FA all individual-level information remains protected behind the normal security mechanisms of the local host system, and the data owner retains full control over the minimum level of aggregation required to allow data to be viewed by the analyst, with protection against unauthorized use (Figure 1). From a statistical perspective, FA allows bi-directional exchange of information between a statistical model and sensitive individual-level patient data via anonymized group-level summary results, without sharing information on individuals. Thus, the data owners retain full local control over security settings that determine the level of aggregation required to protect sensitive personal information.

Whenever multiple raw datasets are used for a study, there is a need to create common study variables, a “Common Data Model”, with common structure and format for the variables. This is a prerequisite for FA as well as for any other strategy for combined analysis of individual level data from different data sources. It is essential that the variables have similar definitions by harmonisation using transparent and well documented algorithms.

A two-step meta-analysis is another alternative to direct access to combined individual level data. With harmonised data at each data node, and a distributed code for analysis, the results from each data node can then be aggregated using conventional meta-analysis. In many instances this will generate results similar to a one-step analysis on individual level data. One potential limitation is that parameters for covariates may be inconsistently estimated across different data sets, but this is not necessarily a concern.

In principle, FA and two-step meta-analysis are statistically equivalent. The main advantage is that FA allows real-time, interactive, centralized, standardized analysis of distributed data, that is, without having to ask data owners to perform specific analysis and provide the results. The main disadvantage is that the required IT infrastructure is considerably more complex for FA than for two-step meta-analyses. While meta-analysis can be performed on standard statistical software and personal computers, FA requires a complex software stack on a federated client–server architecture.
HOW CAN FEDERATED ANALYSES FACILITATE POST-MARKETING STUDIES FROM A REGULATORY PERSPECTIVE?

Characterisation of drug safety profiles is one of the core missions of regulatory agencies. Data available at the time of a marketing authorisation is often not sufficient to fully characterise safety and non-interventional post-authorisation safety studies (PASS) using existing health-care databases are often required. A current example is the characterisation of myocarditis as an adverse reaction from mRNA COVID-19 vaccines.

To assess the practical extent of insufficient sample size in PASS, and the potential of FA to address this concern, we reviewed regulatory assessment reports of PASS final results on the agenda for the plenary meetings of the European Medicines Agency Pharmacovigilance Risk Assessment Committee (PRAC) in 2019. Results from inferential studies involving modelling of covariates were selected for review. The review was restricted to 27 non-interventional PASS with an inferential design and study objective motivating regression analysis (Figure 2). The time periods observed in these studies were on average 9.3 years, ranging from 5 to 19 years. Despite this, 74% (20/27) of the assessments reported issues with poor precision of estimates.

A total of 20 studies were performed using a single data source, but in 70% (14/20) of these we still assessed a FA approach applicable, based on availability of other similarly structured data sources that could have been used. In one of seven studies based on multiple data sources a FA approach had been used, and four studies were considered potentially suitable since multiple similarly structured data sources were used in the study but without pooling of individual level data. This suggests that FA could be applicable in the majority of studies involving multiple datasets.

WHICH STATISTICAL ANALYSES ARE AVAILABLE AND WHAT ARE THE LIMITATIONS?

In our review of PASS the most common statistical methods used were Cox proportional hazards regression (59%; 16/27) and logistic regression (15%; 4/27). Negative binomial regression, Poisson regression, generalised estimation equations (GEE), LASSO regression, and regularised regression were applied in only a few studies. Models incorporating random effects were used in 22% (6/27) of these studies. Propensity scores (PS) were used in 44% (12/27) of the studies. The extent and potential consequences of missing data was reviewed in only 52% (14/27) of the assessment reports. In 48% (13/27) of these studies missing data was stated as a potential concern. The methods used to handle missing data were single imputation (4/13), multiple imputation (3/13), missing category (2/13), complete case analysis (1/13), no method (1/13), or not stated (2/13).

The following sections discuss opportunities and limitations regarding statistical methods and related issues in relation to the findings in our review of PASS study results (Supplemental online-only material: Report Federated analyses – Statistical methods).

Horizontally or vertically partitioned data

A standard FA requires that the data is horizontally partitioned, meaning that all data sources include different sets of patients, but with the same variables. This is also the key type of data pooling needed to increase the sample size and precision of estimates. In vertically partitioned data, on the other hand, all data nodes include the same set of patients, but with different sets of variables. This type of data pooling

---

FIGURE 1 A schematic representation of the concept of a Federated analysis, here illustrated by four separate data sources. The red boxes denote each data owner’s normal security mechanisms for protection of sensitive personal information. The analysis task is submitted to each data server and executed locally. All summary results are then sent back to the analysis client, where they are integrated to a single result. This process may be iterated, and results optimized until the analysis has converged to a final result.
is also of interest, but FA on vertically partitioned data requires the involvement of a trusted third-party facility.

### 4.2 Generalized linear models

Generalized linear models (GLMs) include many common regression models, for example, linear regression, logistic regression, or Poisson regression, which have a broad range of applications in epidemiology. They can be used in FA in a form that gives identical results to the standard formulation of GLMs. GLMs are implemented, for example, in the R/DataSHIELD package for FA.

### 4.3 Time-to-event analyses

The equivalence of the Cox proportional hazards model in a FA compared to a standard analysis has been demonstrated. This approach requires that distinct event times are shared between sites, which potentially could identify a patient. There are also limitations in handling high-dimensional data.

A federated meta-analysis approach using the Cox model has recently been added to the R/DataSHIELD package performing the Cox model separately in each federated data set. The results can then be combined using meta-analysis. In case of rare events the result from such an analysis can be biased because of the normal approximation of the likelihood function. Other options for the likelihood function approximation exist but are not yet implemented as FA but for use in a traditional meta-analysis approach (e.g., the package R/EvidenceSynthesis).

The only option to analyse time to event data federated on individual data, without sharing event time between sites and without using the meta-analysis approach, is to approximate the Cox proportional hazard model with a GLM using Poisson regression.

### 4.4 Propensity scores

Propensity scores based on measured baseline covariates are commonly used to control for confounding. They can be calculated using logistic-binomial regression, which is available as FA. In a FA the propensity scores can either be site-specific, where a model is fitted separately in each site with the possibility of including site-specific measurement covariates to control within-site confounding or fitted using harmonized variables and all observations as if pooled to control between-site confounding. The propensity score can then be used for stratification, reweighting or adjustment, but matching across sites will not be possible.

### 4.5 Missing data

When a complete case analysis is inappropriate, trivial methods for imputation of missing values, such as by imputation of the mean, can be applied. If data is missing-at-random (MAR), maximum likelihood estimation (without imputation) or multiple imputation are recommended. Multiple imputation by Chained Equations (MICE) can be used in FA and is under development to be integrated in the DataSHIELD software. If data is missing-not-at-random (MNAR), model-based imputation is required, but has not been applied in FA to our knowledge. When a variable is missing completely at one or more data sites a possible solution is to replace the covariate based on a prediction model (including uncertainty) using other correlated variables from other data sources. This is equally applicable for physically merged data and FA.

### 4.6 Limitations

Statistical analyses that require the combined or joint distribution of individual-level data from multiple data sources, or are based on non-parametric empirical distributions, which cannot be approximated by parametric distributions, will not be possible or are technically
challenging. Examples of such analyses are the identification of duplicated patient records across datasets, the calculation of the median across multiple datasets, or the analysis of variables with non-parametric statistical distributions whose properties cannot be adequately summarized by a parametric distribution. These issues can be solved by

- the use of alternative statistical measures, for example, by calculating the (weighted) mean instead of the median, or the standard deviation instead of range,
- the use of alternative distributions, for example, deriving the median from a fitted parametric distribution,
- or the transformation of a variable, for example, from a complex, continuous distribution to a categorical distribution.
- the comparison of encrypted patient-identifying features.

4.7 | Computation time

In an FA all participating data servers must be up and running simultaneously for the duration of the analysis. Computation time for a FA is determined by the network latency or lowest specification hardware in the overall system, and will also depend on the number of steps in iterative statistical analyses. The statistical models may have to be re-formulated in a way that requires the computation of additional parameters. For example, a log-Poisson Generalized Linear Model requires additional parameters compared to a Cox proportional hazards model, and therefore increased computation time. Additional processing may also be required to apply data disclosure controls.

5 | WHAT TO CONSIDER REGARDING IT ARCHITECTURE?

Federated analysis requires a complex software stack of applications to coordinate data management and data analysis. One example is the open-source OBIba software application suite developed in the Maelstrom Research project. This aims to facilitate epidemiological research using multiple, physically separate data sources. This involves collecting and harmonising data from different databases, publishing general information about the content of data sources, and creating tools for FA. There is a strong collaboration with the Data to Knowledge Research Group (D2K) at Newcastle University, United Kingdom, which is spearheading the development of DataShield, the key software for FA within the R software for statistical computing. All software developed in Maelstrom Research has a GPL v3 open-source licence. The Maelstrom Research project software was used in the Swedish Medical Products Agency development project as an example because it is rooted in a non-profit organisation, uses open-source code, and is used by academic research groups in Sweden. There are several other technical solutions for FA, but it was not within the scope of this project to compare different software.

Federated analysis requires a basic IT setup with a central access server with R Studio software, which connects to several DataSHIELD servers via secure permanent VPN channels. The client (i.e., analyst) computer has a secure connection to this central access server. Once this central node server is setup at one of the locations, the access to all other servers is through this node server. The implementation and maintenance of such a system requires a strong commitment from the IT support teams at all locations. Integration with existing local IT architecture and IT policies may be challenging (Supplemental online-only material: Report Federated analyses - IT architecture).

6 | WHAT LEGAL IMPLICATIONS SHOULD BE CONSIDERED?

The main advantage of FA is that the processing happens at the node level, which gives more control over sensitive personal data to the local research principals at each node. While the General Data Protection Regulation (GDPR) is a common legislation for all EU countries, national legislation for ethical approval of research may still result in differences between EU countries in views on access to sensitive personal information for research. When FA is used, the local research principal at each node performs several specific processing operations on personal data in the stage prior to the actual FA. The statistical analysis performed locally at the node constitutes a personal data processing operation. The numerical estimates delivered to the central server by each node are not to be construed as personal data when data only consists of aggregated results from statistical calculations.

Responsibility for personal data is based on the real influence an actor has over each processing operation. The term controller means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purpose and means of the processing of personal data. The term processor means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller. There may be multiple controllers for the various processing operations performed. In FA an actor may have a real influence over a given processing operation even if no individual level data has been transferred to that actor. It is therefore appropriate to establish early in the planning phase if there will be one or more research projects and distribute roles and responsibilities accordingly. All agreements should be in writing and regulate roles and responsibilities, as well as liability for any failure to comply with the terms of the agreement.

In traditional register-based research, the role of the controller for different processing activities tends to be assigned based on where the personal data is being processed. This means the role of controller may be transferred between organisations and research principals together with the personal data. When FA is used, responsibility for processing may be divided among the participating research principals, even if the personal data in question is not transferred. It is crucial to clarify procedures, roles, and relationships to determine who is a controller. All collaborating parties should be aware of the division of responsibilities to avoid any party being able to influence the purpose...
or means of processing, thus altering the de facto control of processing. The roles of controller and processor are distributed based on how responsibility for the purpose and means of processing is to be allocated. This in turn depends on the intended nature of collaboration and the influence that each of the participating research principals has on the analysis. Although the primary responsibility is to the data subject, there is also a responsibility to the other parties involved in the project.

We propose three different theoretical models for the distribution of responsibility as controller of personal data in FA (Figure 3). They are intended to facilitate an analysis where actors can determine the purpose and means of personal data processing. It is important that policymakers, researchers, technicians, lawyers, and others with key roles in a research project are sufficiently familiar with how FA works, how data protection regulations should be applied, and the intention behind distributing responsibility in the project. If a processing operation serves common scientific interests, the operators jointly determine the purposes of processing, even if they each have their own specific purposes at an earlier or later stage.

The question of who is to be considered the controller of the personal data being processed is important, given that the controller shall ensure compliance with GDPR in all processing operations for which they are responsible. This implies that the controller shall ensure that data subjects are informed about how their personal data is being processed. Data subjects have the right to obtain the rectification or erasure of personal data concerning them. The controller or processor is also liable to compensate any person who has suffered damage.

There is no requirement pursuant to GDPR for a written agreement between joint controllers, but it is strongly recommended that the division of responsibility, including economic undertakings, and limitations on liability are always documented in a contract or other legally binding document. The division of responsibility and how processing is to be performed within the project must be unambiguous and transparent.

### MODEL 1: Data processing agreements

The research principal for Node 1 is the controller of all personal data processed in the federated analysis. The research principals for Nodes 2 and 3 are processors.

### MODEL 2: Sole overall responsibility

A single research principal with overall responsibility, in combination with nodes with independent responsibility.

### MODEL 3: Joint controllers

Joint overall responsibility in combination with nodes with independent responsibility.

**FIGURE 3** Three theoretical models for the division of controller legal responsibility in a research project that uses Federated analysis. The models illustrate the influence exerted by one or more research principals over operations performed in the central server and the nodes. The models assume that the operations described can be deemed to fall within the scope of the General Data Protection Regulation (GDPR); for example, in terms of responsibility for implementing security measures or technical measures.

## 7 | CONCLUSIONS

Federated analysis allows real-time, interactive, centralized statistical analyses on individual-level data, without actual transfer of sensitive personal data between institutions and countries. It has the potential to facilitate collaboration and improve the precision of postmarketing safety studies, by increasing the quantity, variety, and availability of data needed to study rare adverse events. Our review of post-authorisation studies indicates that lack of precision in such studies is a common limitation. The technique may be particularly attractive for situations where repeated collaborative projects are anticipated, and the cohorts are dynamic. A recent example is the need to characterise timely postmarketing safety of COVID-19 vaccines. The Nordic countries, having very similar nationwide health data resources, would be an attractive area for such collaborations. Rheumatic diseases can serve as another example of a therapeutic area that has seen a rapid development of new medicinal products and a need to further characterise their safety profile in clinical practice, and where there are existing disease registers with similar structures in several different countries. Such situations would likely benefit from using FA.

Federated analysis does not remove the data harmonisation step, requires reliable support for integrating the FA-specific IT architecture with the respective organisation’s general IT architecture and security solutions, and should be based on clear and detailed agreements between involved parties to fully comply with legal obligations. Common statistical methods are available as
mathematical models and software implementations for FA. The implementation of FA requires competent and active involvement of all collaborating parties.

The four full reports from the Swedish Medical Products Agency on IT architecture, legal considerations, statistical methods, and a tutorial are provided in the online Appendix.

**AUTHOR CONTRIBUTIONS**

Rolf Gedeborg conceptualised and drafted the initial version of the manuscript. All other authors contributed with critical review of the manuscript.

**FUNDING INFORMATION**

This research was conducted as a development project by the Medical Products Agency, which is a Swedish Government agency. The study did not receive any external funding.

**CONFLICT OF INTEREST**

Dr Ljung reported receiving grants from Sanofi Aventis paid to his institution outside the submitted work; and receiving personal fees from Pfizer outside the submitted work. All other authors declare no conflict of interest.

**DATA AVAILABILITY STATEMENT**

The data that support the findings of this study are available from the corresponding author upon reasonable request.

**ORCID**

Rolf Gedeborg [https://orcid.org/0000-0002-8850-7863](https://orcid.org/0000-0002-8850-7863)

**REFERENCES**


29. Rassen JA, Avorn J, Schneeweiss S. Multivariate-adjusted pharmacoepidemiologic analyses of confidential information pooled from


**SUPPORTING INFORMATION**

Additional supporting information can be found online in the Supporting Information section at the end of this article.

---