Evaluation of statistical software for federated analysis of multi-site real world studies
Fiona Grimson (fiona.grimson@iqvia.com), Nicolas Niklas, Ruben Hermans, Lorena Cimeanu, Benedikt Maissenhaelter, Joseph Kim

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

Key Takeaways

Federated Analysis (FA) is an alternative to pooling individual patient data (IPD) or meta-analysis that:
- Allows analytical results that are equivalent to pooled IPD
- Fully preserves data privacy

There is full preservation of data privacy, as identifiable individual patient data (IPD) or meta-analysis results can be shared between contributing sites or subgroups as required.

### 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
- 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

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.

### References

[1-21] Refer to the cited literature for further details.

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 PV) (Partitioned Variational Inference) 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 combining 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.

Copyright © 2019-2020 IQVIA and licensor. All rights reserved.

### Keys to Success

- **Data Preparation**: Ensure data is appropriately cleaned and formatted for analysis.
- **Data Privacy**: Implement robust encryption and access controls to protect data at rest and in transit.
- **Data Validation**: Conduct thorough validation checks to ensure data integrity.
- **Data Sharing**: Establish clear guidelines and agreements for data sharing.
- **Data Analysis**: Use advanced statistical methods to derive meaningful insights.
- **Data Interpretation**: Ensure findings are communicated effectively to stakeholders.

### Challenges to Overcome

- **Data Accessibility**: Ensuring data from multiple sources is available and accessible.
- **Data Quality**: Addressing issues of data completeness and accuracy.
- **Data Security**: Protecting data from unauthorized access and breaches.
- **Data Interpretation**: Translating complex results into actionable insights.
- **Data Compliance**: Meeting regulatory requirements for data handling.

### Tools and Technologies

- **Data Management Platforms**: Tools like DataSHIELD for secure data sharing.
- **Statistical Software Packages**: Software like R and Python for statistical analysis.
- **Cloud Computing Services**: Services like Amazon Web Services (AWS) for scalable data processing.
- **Data Governance Solutions**: Solutions for ensuring data adherence to policies and regulations.

### Conclusion

Federated analysis is a promising approach for leveraging multi-site real world datasets while respecting data privacy. Further research and development are needed to expand the range of analytical methods that can be effectively implemented in a federated setting.