

Leveraging FeatureExtraction for Scalable Evidence Generation in a Federated Data Network of Registries

Sarah Gasman¹, Catherine Yount², Anna Sheahan³

1. Johnson & Johnson, La Jolla, California, USA
2. Johnson & Johnson, Raritan, New Jersey, USA
3. Johnson & Johnson, Horsham, Pennsylvania, USA

Background:

Federated Data Networks (FDNs) are a valuable source of real-world evidence, especially for rare diseases like systemic lupus erythematosus (SLE) [1,2]. However, generating reliable evidence in a federated environment comes with unique challenges. Typically, evidence generation is an iterative process, where research questions are refined as results are reviewed. In a federated setting, scripts because sites must run analytic scripts individually and share results, each iteration requires coordination amongst all stakeholders, which can take significant time. It is therefore important to ensure output is comprehensive to reduce the number of iterations needed.

This requires prospectively identifying all potentially relevant cohorts, covariates, supporting data, quality-related metrics, and combinations of these. This shifts the analytic burden upstream and requires researchers to pre-specify the full answer space before they have data to review. For example, important metrics like sub-cohort longitudinal sample size may be readily accessible for a single dataset and can therefore be incorporated into research protocol planning. In federated settings, researchers do not have access to row-level data at other sites, so these metrics are unavailable until after analytic scripts are run. Proactive contingency planning, such as generating alternative sub-cohorts of interest or multiple versions of a cohort definition allow for a quicker pivot and fewer unproductive script runs.

The FeatureExtraction package in the OHDSI HADES suite provides a scalable, standardized framework for efficiently characterizing cohorts [3]. We implemented the FeatureExtraction package as part of the analytics pipeline in a federated data network of SLE registries to generate a comprehensive suite of descriptive analyses and data validation metrics. This approach supported multiple abstracts from this network [4-8].

Methods:

Within a federated network of five SLE registries mapped to the OMOP-CDM and representing over 10,000 patients across 29 countries, we implemented the FeatureExtraction package as part of our analytics pipeline to address challenges around limited iterability. To prespecify the full answer space, we defined 808 distinct cohorts based on clinical and patient characteristics, such as disease activity severity, and damage, treatment exposure, time-from-diagnosis, and combinations of these strata. Additionally, we created 43 custom analyses (**Table 1**) producing 110 custom covariates and included several prespecified FeatureExtraction analyses (such as demographic covariates). Cohorts and analyses covered domains such as demographics, conditions, measurements, and drug exposures. Because there was limited ability to request follow up quality control measures, we also developed cohorts and

covariates that produced these metrics as part of the initial execution. Though all outputs are aggregated to preserve privacy, we disaggregated composite measures when possible, to validate results (for example, creating feature-level covariates to validate domain-level covariates).

Results:

The modular, combinatorial output of the FeatureExtraction package is well-suited to address the constraint of limited iterability in a federated data network. All available covariates were generated across the 808 prespecified cohorts, providing a comprehensive characterization of the network and enabling granular characterization of cohorts of interest.

Creating custom cohorts and covariates resulted in improved harmonization across the registries. For example, four registries completed a Physician Global Assessment (PGA) using a visual analog scale. While three of the four recorded PGA using a scale of 0 to 3, one registry used a scale of 0 to 100. For our custom covariates using PGA score, we harmonized outputs to a 0 to 3 scale so results could be both pooled and easily compared across registries.

FeatureExtraction enabled increased confidence in data outputs and improved interpretability of results. For example, all registries experience significant loss to follow up, though follow up duration varied considerably. For longitudinal analyses, we needed to ensure we were evaluating a consistent group of patients over a comparable time period. We created cohorts of patients with at least 3 visits between registry entry and year 3 of follow-up, harmonizing cohort start date to registry entry date and cohort end date to the date of a patient's year 3 visit. Cohort duration was then consistent between patients and registries, enabling longitudinal analyses using custom covariates assessing metrics like damage accrual and treatment changes over follow up.

Proactively generating multiple versions of the datasets of interest in one script execution allowed us to reduce the number of iterations required to arrive at validated results. Since patients can be exposed to multiple treatments at once, we created cohorts and covariates of mutually exclusive groups of patients based on combinations of treatment classes exposures (e.g. antimalarials and steroids, immunosuppressants and steroids). By cross-referencing cohorts against covariates and checking resulting total patient counts against cohort denominators, we could validate results and more easily identify issues with analytic logic or data quality. For example, we learned we needed to add logic to account for the small number of patients not receiving any treatment for SLE.

The analytic infrastructure described here supported multiple published analyses from this network including studies of disease activity, damage progression, and treatment utilization [1–4]. The comprehensive, pre-specified output enabled generation of insights across disease activity, damage progression, and treatment utilization, including subgroup analyses that emerged after initial results were reviewed, without requiring additional site coordination (**Figure 1**).

Conclusion:

In a federated network setting where ability to iterate is limited, the ability to generate comprehensive results across relevant cohorts in one script execution is a methodological advantage. The

FeatureExtraction package provides a scalable, standardized framework that effectively shifts the analytic burden upstream, maximizing the value of each network execution. This approach is generalizable to other federated networks using the OMOP CDM where iterability is constrained.

Figure 1:

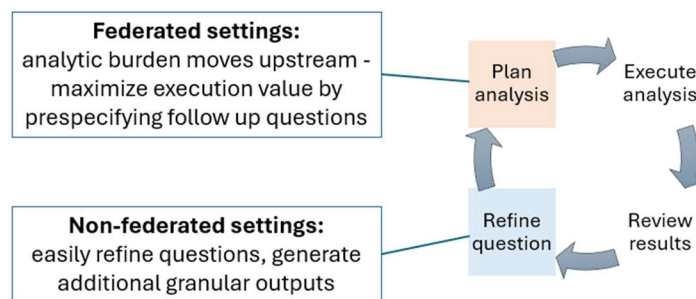
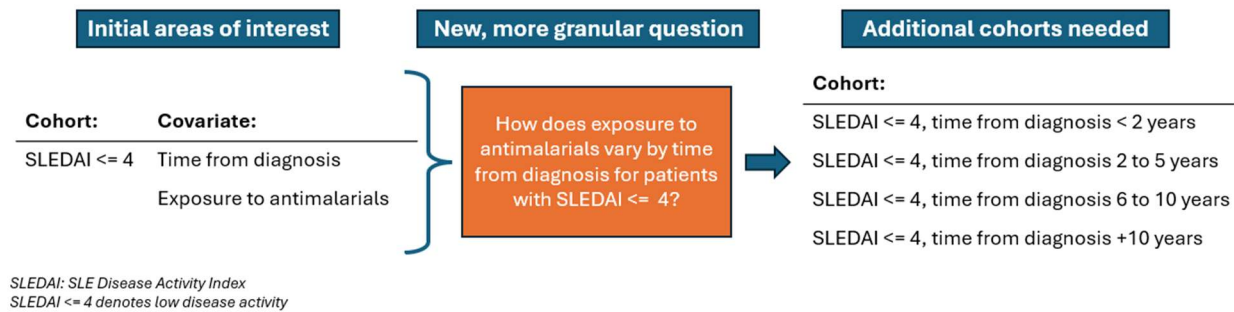


Table 1:

LupusNet: Custom Analyses
BILD Damage Category
BILD Domain Damage
BILD Feature Damage
BILD Damage Score (by visit)
BILD Damage Score (registration and endpoint)
Flare
Heart and Lung Involvement
Lupus Nephritis
Percent of visits with High Disease Activity
PGA Category
PGA score (by visit)
PGA score (registration and endpoint)
SDI Domain Damage
SDI Damage category
SDI Damage Score (by visit)
SDI Damage Score (registration and endpoint)

SDI Feature Damage
SDI Final Damage Score
SLAQ Activity Score (by visit)
SLAQ Activity Score (registration and endpoint)
SLAQ Activity Domain
SLAQ Activity Feature
SLEDAI Activity Score (by visit)
SLEDAI Activity Score (registration and endpoint)
SLEDAI Activity Score (Mean during follow up)
SLEDAI Activity Domain
SLEDAI Activity Feature
Serology Categorical Values
Serology Numeric Values
Severe SLEDAI Feature Presence
Steroid Dosage Categorical Values
Steroid Dosage Numeric Values
Time from Diagnosis, Years
Time from Diagnosis, Categorical
Treatment Class
Treatment Ingredient
Treatment Count
Visits – time between visit dates

References:

1. Blacketer, C., Schuemie, M. J., Moinat, M., Voss, E. A., et al. (2025). Advancing real-world evidence through a federated health data network (EHDEN): Descriptive study. *Journal of Medical Internet Research*, 27, e74119.
2. Blacketer, C. (2025, October). Building OHDSI evidence network [Conference presentation]. Observational Health Data Sciences and Informatics (OHDSI). https://www.ohdsi.org/wp-content/uploads/2025/10/001-Blacketer-Building_OHDSI_Evidence_Network-Clair-Blacketer.pdf
3. OHDSI. (n.d.). FeatureExtraction [Computer software]. GitHub. <https://github.com/OHDSI/FeatureExtraction>
4. Ugarte-Gil MF, Gasman S, Zazzetti F, Orillion A, Sheahan A, Blacketer C, et al. Regional variability in systemic lupus erythematosus damage accumulation by disease activity across the Lupus Federated Data Network (LupusNet). Poster presented at: ACR Convergence; 2025.
5. Zazzetti F, Orillion A, Sheahan A, Blacketer C, Van Speybroeck M, Gasman S, et al. From onset to insights: longitudinal assessment of disease activity, flares, and damage accrual in patients with systemic lupus erythematosus across five registries in the Lupus Federated Data Network (LupusNet). *Ann Rheum Dis*. 2025;84(Suppl 1). Presented at: EULAR; 2025.

6. Orillion A, Zazzetti F, Sheahan A, Blacketer C, Van Speybroeck M, Gasman S, et al. Customized therapy for systemic lupus erythematosus: how disease severity influences the use of corticosteroids and biologics in patients with SLE in the Lupus Federated Data Network (LupusNet) and a US claims database. Poster presented at: ACR Convergence; 2025.
7. Pons-Estel GA, Gasman S, Zazzetti F, Orillion A, Sheahan A, Blacketer C, et al. Systemic treatment patterns in systemic lupus erythematosus: associations with disease severity and organ-specific activity across five registries in the Lupus Federated Data Network (LupusNet). *Ann Rheum Dis.* 2026;85(Suppl 1):POS1055. Presented at: EULAR; 2026.
8. Michaud K, Gasman S, Zazzetti F, Orillion A, Sheahan A, Blacketer C, et al. Age at onset and its association with damage accrual in systemic lupus erythematosus: insights from five registries in the Lupus Federated Data Network (LupusNet). *Ann Rheum Dis.* 2026;85(Suppl 1):POS0731. Presented at: EULAR; 2026.