Federated real-world data and common data models

And how they enable integrated analyses of subject-level data from large data sources while preserving patient privacy

By Rima Izem, PSI RWD SIG

Federated analyses allow data integration from multiple sources using statistical methods that preserve patient privacy. Recently, federated healthcare networks like the US Sentinel System and the European DARWIN and EHDEN networks have made significant progress. These networks enable various data custodians to collaborate on public health analyses without sharing individual-level data.

A key achievement in these networks is mapping each data source to a common data model (CDM). This model reduces between source heterogeneity in data-collection, thereby facilitating integrated analyses. The OMOP CDM and the US Sentinel CDM are widely used, organizing data into subject-level and healthcare encounter-level categories with naming conventions and definitions for commonly used variables in integrated analyses. Contributors to federated networks convert their data to a CDM, maintaining their original data structures. This conversion process is labor-intensive and often funded by end-users of the integrated analysis. Once data sources are mapped to the CDM, consistent methods are applied across sources, allowing for reliable summary-level data integration.

Federated analyses require new statistical methods and tools. For example, traditional software implementing parametric or non-parametric statistical modeling assume centralized data access to subject-level data, whereas federated data is distributed. This necessitates rethinking the analyses performed behind the firewall of each data source, and the summary output from each source, needed for reliable data integration.

Federated analyses represent a growing area of collaboration in observational health research, and the use of digital health technology, offering a robust framework for integrating diverse healthcare data while maintaining patient privacy.