dsOMOP: Federated Analysis of Harmonized Clinical Data Combining OMOP CDM and DataSHIELD in a DATOS-CAT Cohort Use Case

David Sarrat-González[1], Judith Martinez-Gonzalez[2,3], Xavier Escribà-Montagut[1], Aikaterini Lymperidou[3,4], Ramón Mateo[1,3], Rafael de Cid[4], Juan R González[1] and Alberto Labarga[2]

[1]Barcelona Institute for Global Health (ISGlobal). [2]Barcelona Supercomputing Center (BSC). [3]Institute for Bioengineering of Catalonia (IBEC).
[4]Genomes for Life - GCAT Lab - Germans Trias i Pujol Research Institute (IGTP).

Introduction

The **DATOS-CAT project** aims to enhance the visibility and scientific impact of population-based cohorts developed in Catalonia, such as GCAT | Genomes for Life and the COVICAT-CONTENT subcohort. These cohorts provide valuable data for research, but the variability and sensitive nature of the data pose significant challenges, especially in terms of data sharing between institutions. To address these issues, the project has adopted the **Observational Medical Outcomes Partnership Common Data Model (OMOP CDM)** to enable **standardized analysis** and **DataSHIELD** technology to perform **federated non-disclosive statistical analyses**.

What is DataSHIELD?

DataSHIELD is a technology that enables the analysis of sensitive health data across different research sites **without physically sharing the data**. It works by allowing statistical analyses to be performed on individual-level data, but these data never leave their original location. Instead, DataSHIELD **transfers the commands to the data** and then **aggregates the results from each site**. This means researchers can collaborate and analyze combined datasets from various cohorts without ever accessing the raw, sensitive data directly. This allows DATOS-CAT to maintain **patient confidentiality** and adhering to **privacy regulations** (such as GDPR) while still enabling a thorough and collaborative analysis of clinical data across its multiple participant sources.

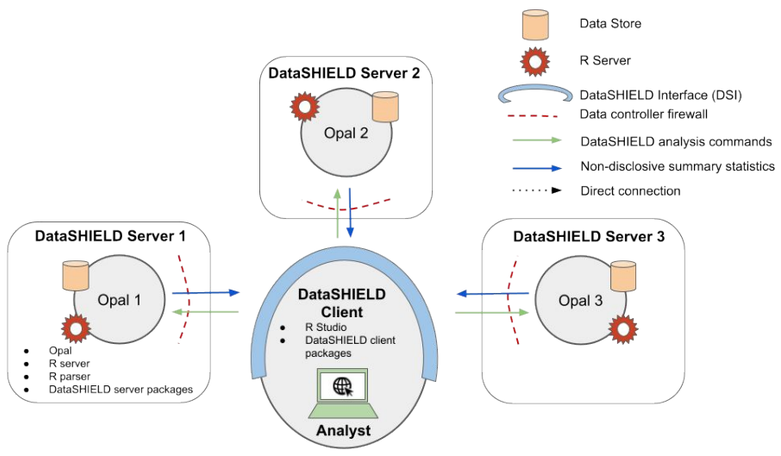

Figure 1: Multi-site DataSHIELD infrastructure architecture using Opal servers

The dsOMOP package

The **dsOMOP package** is designed to facilitate the interaction with remote databases formatted in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) within a DataSHIELD environment. It provides a suite of functions that allow users to **fetch** and **transform data** from these databases into a format that is **intelligible** and **usable** within the DataSHIELD analytical workflow. This direct integration ensures that data analysis complies with the DataSHIELD's **security model** and **disclosure checks** system, which are crucial for maintaining the privacy and security of the data.

Structure

The dsOMOP ecosystem comprises **two essential components** designed to work in tandem: the server-side package (**dsOMOP**) and the client-side package (**dsOMOPClient**).

- **dsOMOP** is installed on the DataSHIELD server by the institution's data manager and is responsible for direct interactions with the OMOP CDM databases. It retrieves, transforms, and integrates the data in a format compatible with DataSHIELD's analytical tools.

  Code available at: **https://github.com/isglobal-brge/dsOMOP**

- **dsOMOPClient** is used locally by researchers and data analysts, it orchestrates the communication between the dsOMOP package on the server and the target OMOP CDM database, allowing for the construction of use-case specific datasets.

  Code available at: **https://github.com/isglobal-brge/dsOMOPClient**

Dataset construction

The dsOMOP package is designed to enable users to orchestrate the **creation of datasets** that meet their **specific research needs**. The package allows users to **define their data queries** with precision, ensuring that only the relevant data is fetched. This process is facilitated by dsOMOP's robust support for database interaction, which includes methods for examining table contents, columns, and concept catalogs. Users can filter data based on their specific criteria, making the data selection process both flexible and efficient.

In addition to data selection, dsOMOP **automatically translates all concepts** present in the database into a format that is **easily understandable** for users. This is achieved by utilizing the concept names from the registered vocabularies within the OMOP CDM. As a result, the retrieved data is presented in a **user-friendly format**, allowing researchers to quickly grasp the content and context of their datasets.

The package also handles the **categorization of data** based on its nature. For instance, it can distinguish between **longitudinal data**, which involves multiple records over time for the same entity, and **relational data**, which links multiple instances across different tables. This automatic categorization ensures that data is always served in the **most appropriate format** according to its nature, making it **readily accessible** for subsequent analysis. All of this is performed without the need for supervision by the researcher.

Support for community extensions

While dsOMOP acts as an interface between DataSHIELD servers and OMOP CDM databases, the **potential for automation** or **streamlining** of processes through the creation of supplementary functions, scripts, and packages is vast.

An example of this approach is **dsOMOPHelper**, a complementary package we have developed alongside dsOMOP. This package **significantly reduces the complexity** of using dsOMOP for most simple use cases, where data from an OMOP CDM database may be used for epidemiological studies within the DataSHIELD environment.

Code available at: **https://github.com/isglobal-brge/dsOMOPHelper**

We **strongly encourage** the community to develop tools that build upon dsOMOP, tailoring them to specific use cases and research needs. Such community-driven development not only enhances the utility of dsOMOPClient but also fosters a collaborative ecosystem around the combined use of both DataSHIELD and OMOP CDM.

Results

Thanks to the development and implementation of the dsOMOP package, we have been enabled to effectively **combine and integrate** the use of various **core components** of the DATOS-CAT project. This integration has allowed us to successfully perform **comprehensive** and **federated analyses** on **harmonized clinical data**.

Specifically, we have been able to:

- **Integrate data** from the various participant cohorts into the DataSHIELD environment in a **standardized format** using the OMOP CDM format.
- **Conduct epidemiological studies** while maintaining strict adherence to **privacy regulations** and ensuring **data confidentiality**.
- Leverage the full suite of **DataSHIELD analytical tools** to perform **non-disclosive statistical analyses** on combined datasets from **multiple sources**.
- Facilitate **data sharing** and **collaboration** between different research centers, enabling a more integrated and **cooperative research environment**.
- Enhance the ability to perform **cross-cohort analyses**, **combining insights** from multiple datasets to generate more robust and comprehensive **research findings**.

The dsOMOP package has proven to be a crucial asset in achieving the project's objectives, enabling **efficient**, **secure**, and **collaborative** data analysis across diverse and sensitive clinical datasets.

Collaborating centers

IBEC · BSC · Clínic Barcelona · cnag centre nacional d'anàlisi genòmica / centro nacional de análisis genómico · CRG Centre for Genomic Regulation · elixir SPAIN · INB Spanish National Bioinformatics Inst · IGTP Germans Trias i Pujol Research Institute · ISGlobal

Funding

Financiado por la Unión Europea NextGenerationEU · Gobierno de España · Ministerio de Ciencia e Innovación · Plan de Recuperación, Transformación y Resiliencia · Next Generation Catalunya · Generalitat de Catalunya

Funded by the "Complementary Plan for Biotechnology Applied to Health," coordinated by the Institute for Bioengineering of Catalonia (IBEC) within the framework of the Recovery, Transformation, and Resilience Plan (C17.I1) - Funded by the European Union - NextGenerationEU