DATOS.CAT: Improving the level of the interoperability of GCAT in line with the FAIR principles with OntoBridge, dsOMOP and OMOP Beacon

Aikaterini Lymperidou[1,2], David Sarrat-González[3], Ramon Mateo-Navarro[2,3], Guillem Bracons-Cucó[4], Aurora Moreno-Racero[2,5], Jordi Rambla de Argila[5], Liina Lagirnaja[5], Carles Hernandez-Ferrer[6], Salvador Capella-Gutierrez[6], Santiago Frid[4], Juan R González[3] and Rafael de Cid[1]

[1]*Genomes for Life- GCAT lab- Germans Trias i Pujol Research Institute, Badalona, Spain*
[2]*Institute for Bioengineering of Catalonia (IBEC), Barcelona, Spain*
[3]*Barcelona Institute for Global Health (ISGlobal), Barcelona, Spain*
[4]*Hospital Clinic de Barcelona, Barcelona, Spain*
[5]*Center for Genomic Regulation, Barcelona, Spain*
[6]*Barcelona Supercomputing Center, Barcelona, Spain*

Keywords
OMOP CDM, ds.OMOP, OMOP Beacon,

1. Introduction

Collecting data over the long term enables researchers to monitor disease progression, uncover patterns related to environmental and genetic risks, and evaluate the effectiveness of various treatment approaches. The GCAT Genomes for life cohort is an initiative in Catalonia that has successfully recruited 20.000 participants between 40 and 65 years and gathers clinical, genetic and lifestyle information. GCAT provides inestimable data that can contribute to our understanding about various diseases. Within the framework of the DATOS.CAT, we developed and tested some procedures for improving the level of interoperability of the GCAT Cohort and consequently of other big cohorts like this in the context of the FAIR data principles (Findable, Accessible, Interoperable, Reusable) to facilitate their scientific use.

2. Methods

In the context of the DATOS.CAT, we suggest a groundbreaking pipeline that can be applicable to population-based cohorts, improving the level of the interoperability of the data in line with the FAIR principles. This procedure includes: the transformation of the local database to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) with OntoBridge, statistical analysis with dsOMOP and the exploration of the data with the tool OMOP Beacon.

First, we introduce OntoBridge, a tool based on ontologies to transform local databases into standards in a way that maintains the semantics data. OntoBridge is based on a system that utilizes three layers of different ontologies. On one hand, it models the data to be transformed from the local database, and on the other hand, it models the standard model to which the transformation is intended. Finally, a third layer semantically and syntactically maps both models to enable the transformation. During the development of the project, a challenge was identified regarding the representation of certain local demographic variables in the OMOP model, due to the absence of specific concept identifiers for these variables.

Beyond data harmonization, an integral part of this initiative involves federated analysis of the standardized data. DataSHIELD is a software primarily used through R and widely adopted across

SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025

Europe. Its purpose is to enable federated data analysis while preserving data privacy. This open-source software has been well received due to its flexibility and ability to promote collaboration among researchers. These features have led to the development of new tools that extend its functionalities, allowing researchers to perform more precise analyses designed to match their specific needs.

One such tool is the dsOMOP package, which integrates the OMOP CDM into the DataSHIELD framework. dsOMOP allows researchers to interact with OMOP CDM-formatted databases while inheriting DataSHIELD's core principles of privacy and secure analysis. The OMOP CDM provides a standardized framework for organizing and transforming local clinical data, enabling seamless data exchange and analysis across different cohorts and research institutions.

The integration of DataSHIELD and dsOMOP within the scope of the project brings additional analytical flexibility and unlocks the potential for large-scale, standardized research across its multiple cohorts and institutions. It ensures that the transformed GCAT data can be analyzed within a trusted environment while retaining all the advantages of standardization and preserving participant privacy.

In addition, the European Genome and Phenome Archive (EGA) infrastructure serves as the foundation for storing genomic data from individuals in Catalonia, providing high-value resources for research into disease development and underlying processes. There are two Beacon networks created for the purpose of that project: One to enable the discovery of genomic features, the EGA Beacon network, following the GA4GH standard, and one to facilitate the discovery of clinical information. The OMOP Beacon is developed to enable federated queries on the GCAT cohort's clinical data, allowing researchers to access and explore this information in a secure, privacy-preserving manner.

The integration of the OMOP CDM and the OMOP Beacon plays a crucial role in ensuring the accessibility and interoperability of clinical data enabling federated queries of the GCAT cohort's clinical data and allowing researchers to access and explore this information in a secure, privacy-preserving manner.

3. Results

The DATOS.CAT pipeline integration enables researchers to directly access vital information sources, fostering a deeper understanding of the relationship between genetic profiles and clinical outcomes. This approach unlocks the potential for groundbreaking discoveries and generates new knowledge with the capacity to significantly transform healthcare practices worldwide. DATOS-CAT aims to enhance the global competitiveness of high populated cohorts like GCAT by linking them to comprehensive genomic and clinical data, ultimately driving greater societal benefits.

Declaration on Generative AI

The authors have not employed any Generative AI tools.