Digital Health and Informatics Innovations for Sustainable Health Care Systems
J. Mantas et al. (Eds.)

doi:10.3233/SHTI240378

DATOS-CAT: OMOP-Common Data Model for the Standardization, Integration and Analysis of Population-Based Biomedical Data in Catalonia

Aikaterini LYMPERIDOU[a,b,1], Judith MARTINEZ-GONZALEZ [a,c], Guillem BRACONS CUCÓ[a,d], Santiago FRID[d], Rafael DE CID[b] and Alberto LABARGA[c]
[a] *Institute for Bioengineering of Catalonia, Barcelona, Spain*
[b] *Genomes for Life-GCAT-Germans Trias i Pujol Research Institute, Badalona, Spain*
[c] *Barcelona Supercomputing Center, Barcelona, Spain*
[d] *Hospital Clinic de Barcelona, Barcelona, Spain*
ORCiD ID: Aikaterini Lymperidou https://orcid.org/0009-0002-3425-2188, Judith Martinez-Gonzalez https://orcid.org/0009-0009-9441-5971, Guillem Bracons Cucó https://orcid.org/0000-0003-1274-7403, Rafael de Cid https://orcid.org/0000-0003-3579-6777, Alberto Labarga https://orcid.org/0000-0001-6781-893X

Abstract. Transforming the population based biomedical cohort into the Common Data Model (OMOP-CDM) empowers researchers to access direct sources of information, enabling a deeper understanding of how genetic profiles relate to clinical outcomes and providing new knowledge that can significantly influence health care practices around the world.

Keywords. OMOP-CDM, FAIR principles, GCAT, DATOS-CAT

1. Introduction

In the context of personalized medicine, long-term data collection allows researchers to track how diseases progress over time, identify patterns of environmental and genetic risk, and assess the impact of different treatment strategies. The GCAT (Genomes for Life) cohort [1] has successfully recruited 20.000 participants in Catalonia gathering lifestyle, health and genetic information. This initiative stands as a cornerstone of biomedical research in Europe, providing invaluable data that can drive forward our understanding of various diseases and contribute to the development of personalized medicine. A key aspect of the DATOS-CAT project is to manage and analyze the data in the context of the **FAIR data principles** (Findable, Accessible, Interoperable, Reusables). The European Genome-Phenome Archive (EGA) serves as the backbone for storing genomic data, while the **Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM)** provides a standardized framework for working with structured clinical data. The **GA4GH Beacon** [2] project and **DATASHIELD** [3] are also key components of this framework, enabling the discovery of data sources that

[1] Corresponding Author: Aikaterini Lymperidou; E-mail: alymperidou@igtp.cat.

meet specific genomic or phenotypic criteria, with **specific developments to analyze OMOP-CDM resources in a federated way**.

2. Methods

Following OHDSI guidelines, we used *WhiteRabbit* for data exploration, *Rabbit-in-a-Hat* for ETL design. *Meltano* was then used for data ingestion and *DBT* for data transformation. The pipeline [4], includes extensive quality control using Great Expectations custom OMOP tests, and final review using OHDSI Data Quality Dashboard and Achilles libraries.

3. Results, Discussion and Conclusions

Firstly, the primary OMOP mapping of nearly 20.000 individuals is completed, with a comprehensive population of all the key OMOP clinical data tables. Later on, the Metabolomic, Proteomic and Environmental datasets will be transformed.

The adoption of OMOP-CDM ensures that heterogeneous datasets can be seamlessly integrated and used in a variety of research initiatives, making federated data analysis also possible. However, one of the main challenges was the standardization of certain concepts, since some of them were not previously defined or lacked clear standards in the literature. Therefore, we suggest that data standardization should start from the initial phase of data collection. It is essential to establish uniform standards from the beginning of the process, which will ensure consistency and quality of data, and facilitate more robust and meaningful future analyses.

DATOS-CAT seeks to make the GCAT cohort more competitive at a global level, by using the OMOP-CDM and connecting the resulting database with direct sources of genomic and clinical information. This enables a deeper understanding of how genetic profiles relate to clinical outcomes, unlocking potential breakthrough discoveries and providing new knowledge that can significantly influence health care practices around the world.

References

[1] Obón-Santacana M, Vilardell M, Carreras A, Duran X, Velasco J, Galván-Femenia I, Alonso T, Puig L, Sumoy L, Duell EJ, Perucho M, Moreno V, de Cid R. GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. BMJ Open. 2018 Mar 27;8(3):e018324. doi: 10.1136/bmjopen-2017-018324. PMID: 29593016; PMCID: PMC5875652.

[2] Rambla J, Baudis M, Ariosa R, Beck T, Fromont LA, Navarro A, Paloots R, Rueda M, Saunders G, Singh B, Spalding JD, Törnroos J, Vasallo C, Veal CD, Brookes AJ. Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. Hum Mutat. 2022 Jun;43(6):791-799. doi: 10.1002/humu.24369. Epub 2022 Apr 8. PMID: 35297548; PMCID: PMC9322265.

[3] Marcon Y, Bishop T, Avraam D, Escriba-Montagut X, Ryser-Welch P, Weather S, Burton P, González JR. Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD. PLoS Comput Biol. 2021 Mar 30;17(3):e1008880. doi: 10.1371/journal.pcbi.1008880. PMID: 33784300; PMCID: PMC8034722.

[4] http://github.com/DATOS-CAT