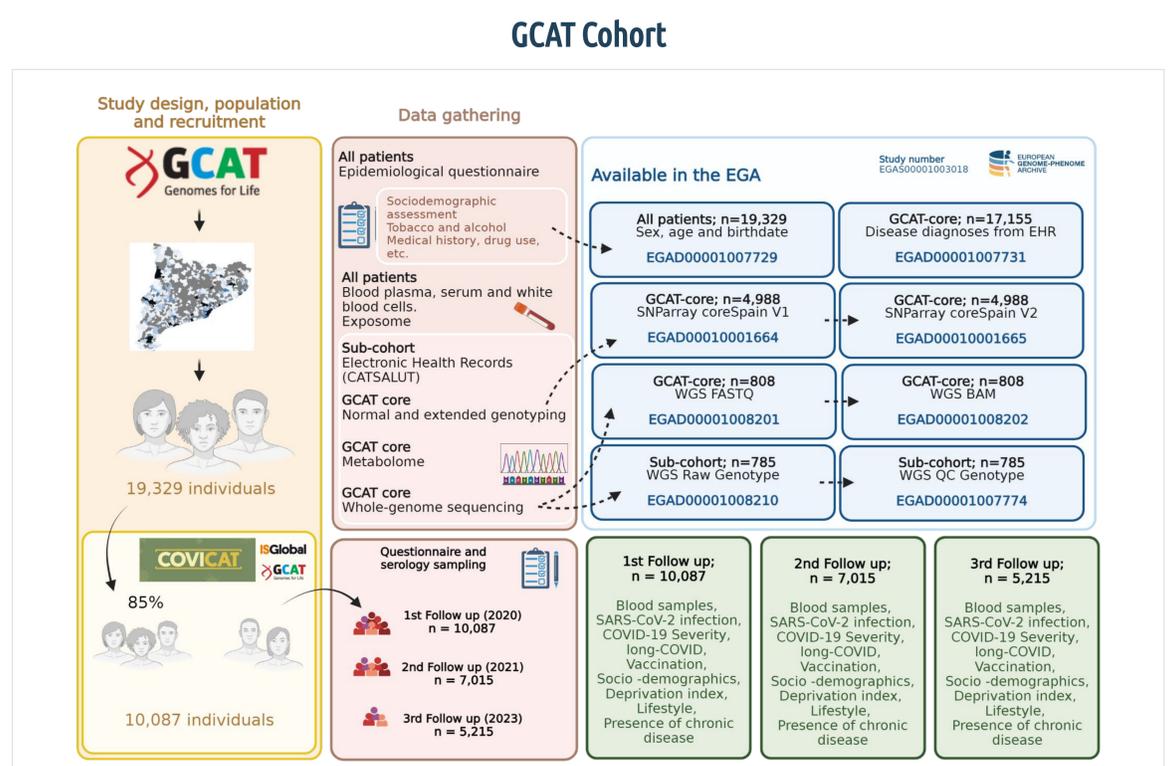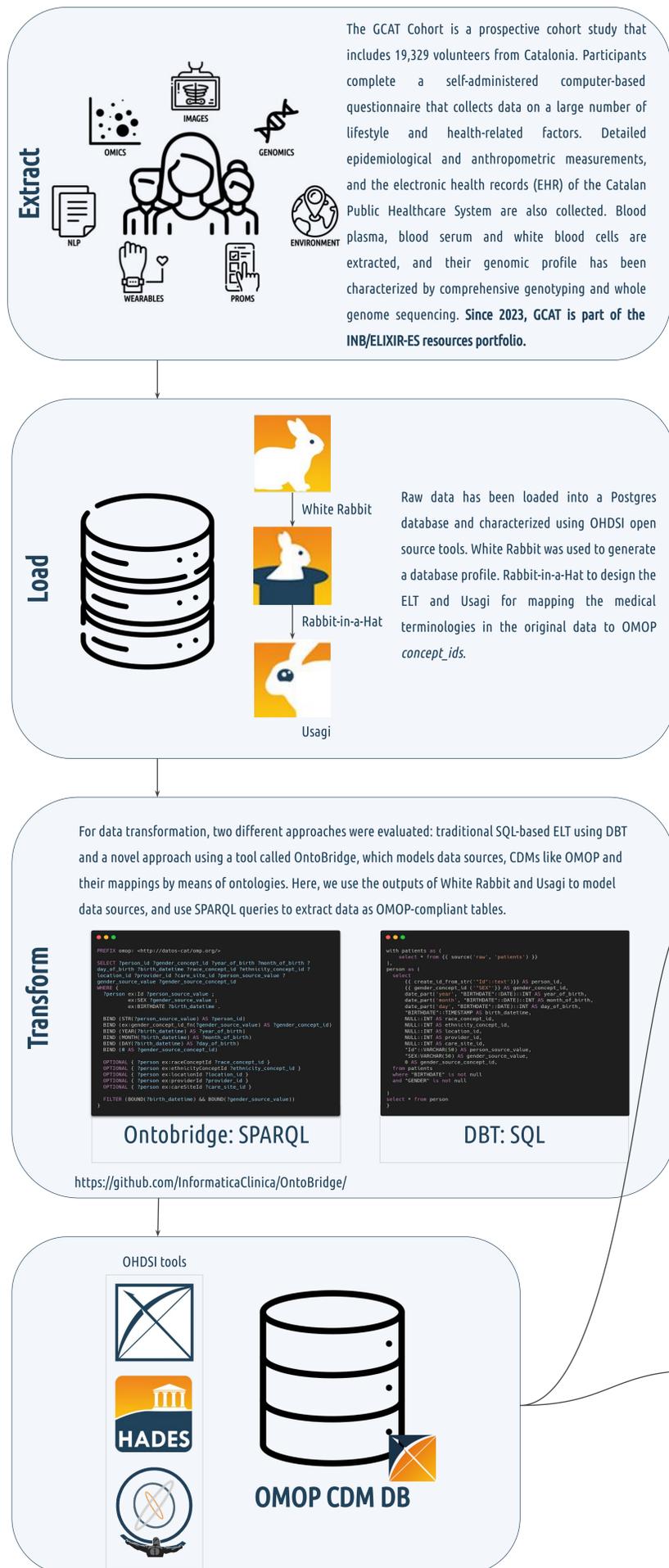# DATOS-CAT: Leveraging OMOP CDM for the standardization, integration and analysis of population-based biomedical data in Catalonia

Judith Martinez-Gonzalez [1,2], David Sarrat-González [3], Ramon Mateo-Navarro [1,3], Guillem Bracons Cucó [1], Aikaterini Lymperidou [1,4,8], Marta Huertas [1,5], Sergi Aguiló-Castillo [2,6], Salvador Capella-Gutiérrez [2,6], Rafael de Cid [4,6,8], Santiago Frid [7], Juan R González [3] and Alberto Labarga [2,6]

[1] Institute for Bioengineering of Catalonia (IBEC). [2] Barcelona Supercomputing Center (BSC). [3] Barcelona Institute for Global Health (ISGlobal). [4] Genomes for Life- GCAT lab- Germans Trias i Pujol Research Institute (IGTP). [5] Centre for Genomic Regulation (CRG). [6] Spanish National Bioinformatics Institute (INB/ELIXIR-ES). [7] Hospital Clínic de Barcelona. [8] Grup de Recerca en Impacte de les Malalties Cròniques i les seves Trajectòries (GRIMTra).

Extract

The GCAT Cohort is a prospective cohort study that includes 19,329 volunteers from Catalonia. Participants complete a self-administered computer-based questionnaire that collects data on a large number of lifestyle and health-related factors. Detailed epidemiological and anthropometric measurements, and the electronic health records (EHR) of the Catalan Public Healthcare System are also collected. Blood plasma, blood serum and white blood cells are extracted, and their genomic profile has been characterized by comprehensive genotyping and whole genome sequencing. **Since 2023, GCAT is part of the INB/ELIXIR-ES resources portfolio.**

Load

Raw data has been loaded into a Postgres database and characterized using OHDSI open source tools. White Rabbit was used to generate a database profile. Rabbit-in-a-Hat to design the ELT and Usagi for mapping the medical terminologies in the original data to OMOP *concept_ids*.

White Rabbit

Rabbit-in-a-Hat

Usagi

Transform

For data transformation, two different approaches were evaluated: traditional SQL-based ELT using DBT and a novel approach using a tool called OntoBridge, which models data sources, CDMs like OMOP and their mappings by means of ontologies. Here, we use the outputs of White Rabbit and Usagi to model data sources, and use SPARQL queries to extract data as OMOP-compliant tables.

Ontobridge: SPARQL

DBT: SQL

https://github.com/InformaticaClinica/OntoBridge/

OHDSI tools

HADES

OMOP CDM DB

GCAT Cohort

Study design, population and recruitment

Data gathering

Federated data analysis: dsOMOP

The **dsOMOP package** is designed to facilitate the interaction with remote databases formatted in the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) from within a DataSHIELD environment. DataSHIELD allows statistical analyses to be performed on individual-level data, but without being transferred from their original location.

```
o$auto(tables = c("measurement"),
       concepts = c(3043111),
       columns = c("value_as_number"))

ds.table("omop$platelet_mean_volume_entitic_volume_in_blood_by_automated_count
.value_as_number")
```

https://github.com/isglobal-brge/dsOMOP

Histogram of study

platelet_mean_volume_entitic_volume_in_blood_by_automated_count.value_as_number
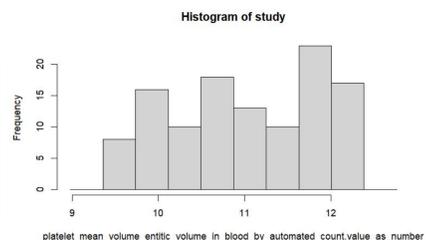
Federated data discovery: G4GH Beacon v2

Beacon

The **GA4GH Beacon v2** enables the federated discovery of data sources that meet specific genomic or phenotypic criteria. We developed a **plugin** that allows integrating **OMOP CDM** databases into the Beacon ecosystem.

https://gitlab.bsc.es/impact-data/impd-beacon_omopcdm

```
curl -s http://omop.beacon.bsc.es/api/individuals | jq
'[.response.resultSets[].results[].measures[] | select(.assayCode.id ==
"LOINC:32623-1") | {id: .assayCode.id, label: .assayCode.label, value:
.measurementValue.value, unit:
.measurementValue.unit.id}]'_mean_volume_entitic_volume_in_blood_by_automate
d_count.value_as_number)

[{
    "id": "LOINC:32623-1",
    "label": "Platelet mean volume [Entitic volume] in Blood by Automated
count",
    "value": 10.8,
    "unit": "UCUM:fL"
  },
  {
    "id": "LOINC:32623-1",
    "label": "Platelet mean volume [Entitic volume] in Blood by Automated
count",
    "value": 9.6,
    "unit": "UCUM:fL"
  }...]
```

[1] Obón-Santacana, M. et al. GCAT|Genomes for life: a prospective cohort study of the genomes of Catalonia. BMJ Open 8, 2018

[2] Rambla J. et al. Beacon v2 and Beacon networks: A "lingua franca" for federated data discovery in biomedical genomics, and beyond. Hum. Mutat., 43, 791–799, 2022.

[3] Marcon, Y. et. al Orchestrating privacy-protected big data analyses of data from different resources with R and DataSHIELD. PLoS computational biology, 17(3), 2021.

Code available at http://github.com/DATOS-CAT