CaRROT-CDM

Appleby, Phil; Masood, Erum; Milligan, Gordon; Macdonald, Calum ; Quinlan, Philip; Cole, Christian

DOI:
10.5281/zenodo.10707025

Publication date:
2023

CaRROT-CDM: An Open-Source Tool for Transforming Data for Federated Discovery in Health Research

Philip Appleby[1], Erum Masood[1], Gordon Milligan[1], Calum Macdonald4, Philip Quinlan[2], Christian Cole[1,3]

[1] *Health Informatics Centre, University of Dundee, Scotland UK*

2 *Digital Research Service, University of Nottingham, UK*

3 *Population Health and Genomics, School of Medicine, University of Dundee, Scotland UK*

4 *Health Data Research, London, UK*

Background

The software described is in use to support federated discovery for health data under the control of numerous organisations. Data query hubs have been established for two major projects so far: the CO-Connect COVID Data project and the Alleviate Pain Data Hub.

Problem Statement

Lack of Data Standardisation. This has resulted in limited ability to compare and link data between various health data research cohorts due to varying:

- Data schema and storage
- Formats (Free text, codes, vocabularies)
- Health data coding standards
- Biometric data (medical conditions and procedures, demographics, and prescribed pain relief medications)

CaRROT CDM Transformation to the OMOP standard

The ETL tool (CaRROT-CDM) is designed to process input data from the whole range of data partners. Data partner data include information gathered for specific studies and collected routinely, this leads to cohorts and datasets of greatly varying sizes.
Data protection concerns mean, that testing is only ever conducted on synthetic data in development compute environments.

CaRROT-CDM Current Work

Use of Python Pandas in earlier versions- imposed memory and speed limits. In the current version data are streamed record by record and OMOP data objects are defined via configuration.

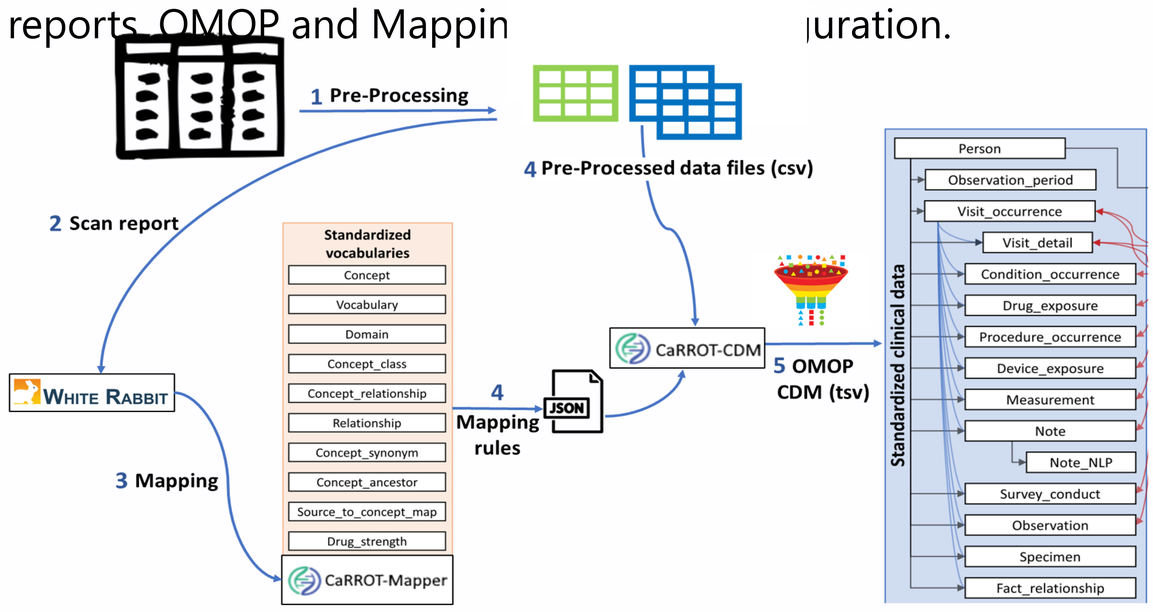
Methods – CaRROT-Mapper and CaRROT-CDM

The Federated Data Hubs are aimed at making data Findable, Accessible, Interoperable and Reusable (FAIR). Data standardisation in Co-Connect and Alleviate is to the ***Observation Medical Outcomes Partnership Common Data Model version 5.3.1*** **(OMOP CDM 5.3.1).**

A Semi-automated data standardisation methodology is used.

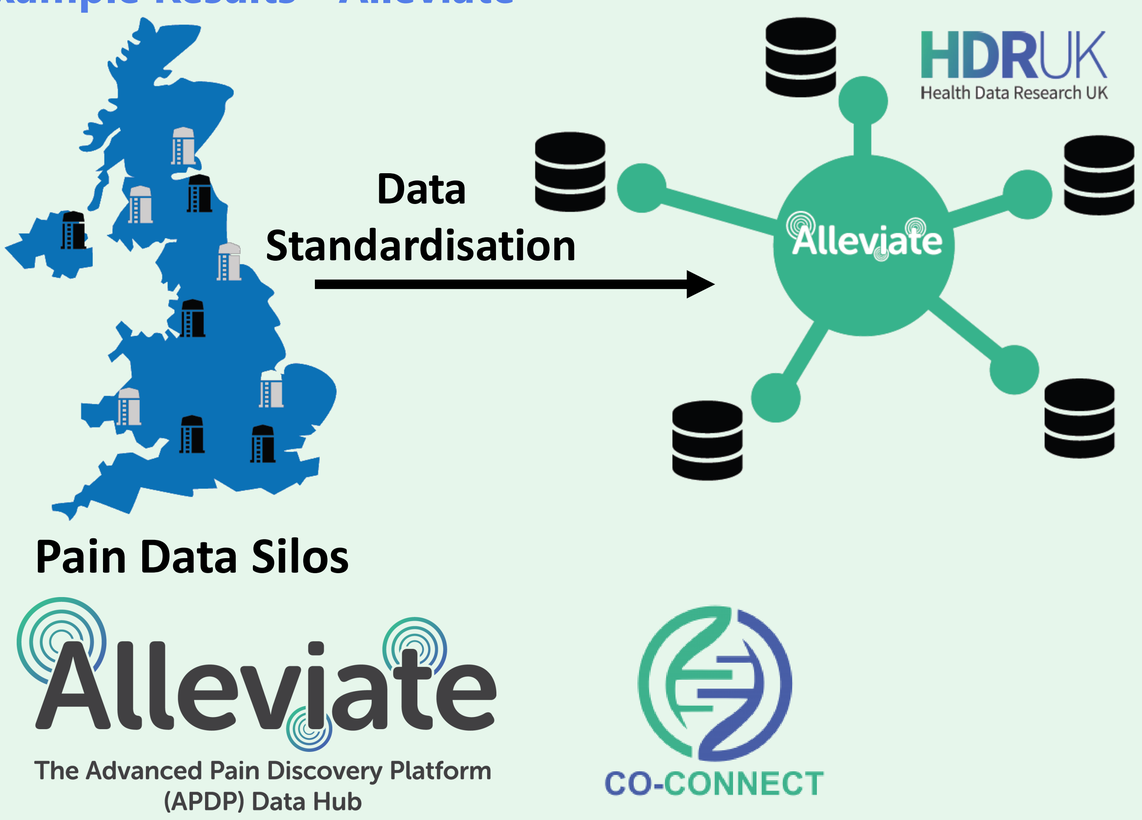
- **Pre-Processing,** standardise dates, units and numeric values, remove duplicates, interpolate missing information (where possible) and pseudonymise personal information.
- **CaRROT Mapper,** a webapp using dataset metadata (produced by the WhiteRabbit data scanning tool) to output term and structural mappings from the standard vocabularies defined in the OMOP Common Data Model.
- **CaRROT CDM,** a python tool installed via "pip install" at Data Partner sites, to perform transformation on data using the mappings generated from CaRROT Mapper. CaRROT-CDM includes commands to generate synthetic data from WhiteRabbit scan reports, OMOP and Mappin uration.

CaRROT Mapper / CaRROT-CDM data flow: Preprocessing, Scanning, Mapping and Transformation.



CaRROT-CDM data transformation – allows for many-to-many relationship between input and output records.

Example Results - Alleviate



Alleviate
The Advanced Pain Discovery Platform (APDP) Data Hub

CO-CONNECT

Data Discovery

- Federated data are standardised to OMOP CDM
- Approved researchers can safely and securely perform federated data queries using the HDR UK Cohort Discovery tool.
- The discovery tool can query multiple datasets from across the UK at the same time to identify suitable data for use in research.
- Following data discovery, researchers apply to data controllers for access to the relevant study data.

Trusted Research Environments

- OMOP Data can be imported into Trusted Research Environments.

Findable
Accessible
Interoperable
Reusable

https://github.com/HDRUK/CaRROT-CDM