Mapping UK Biobank to the OMOP CDM: challenges and solutions

PRESENTER: **Sofia Bazakou**

Background

UK Biobank[1] (UKB) is a large-scale registry containing medical and genetic data from 500,000 consented participants from the UK's general population, aged between 40 and 69 years (Figure 2). UKB is an extraordinary resource for human health research, accessible to approved research initiatives worldwide.

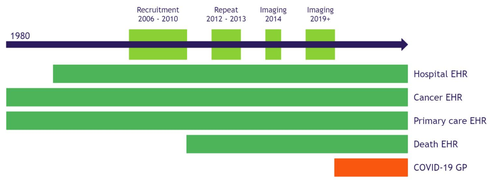

Figure 2. UKB data structure and timeline. The data include multiple baseline assessments (light green), such as surveys, samples, and imaging, linkages to electronic health records (EHR) from different sources (dark green), and information on COVID-19 testing (red). Picture adapted from Prof. Denaxas, UCL.

As part of the European Health Data Evidence Network[2] (EHDEN), The Hyve collaborated with University College London (UCL) to map the UKB data to the OMOP CDM v5.3. The Hyve performed the technical part of the mapping, whilst UCL provided the source data expertise.

The main goal of the collaboration was to make the dataset available for research related to the COVID-19 pandemic.

The UKB data conversion effort came with several challenges:

- Conversion of a large wide format table to long format. For each patient, a wide variety of variables and time points needed to be extracted from a 500,000 by 9,000 table.
- Large heterogeneity of source terms amongst data providers.
- Terms in free text or captured using a mix of ontologies (some of which have now been deprecated).
- Developing the ETL scripts relying entirely on synthetic data from the WhiteRabbit scan report.
- Working with an evolving data source.

Methods

We initially used White Rabbit to generate synthetic datasets. Next, we created the syntactic mappings to the OMOP CDM and the documentation with Rabbit in a Hat (RiaH), and the semantic mappings of source codes to standard concepts using Usagi (free text fields).

A powerful and flexible **ETL framework** together with existing **open-source tools** from the OHDSI suite allowed us to perform the conversion and to deliver a high-coverage mapping **without direct access** to the UKB data.

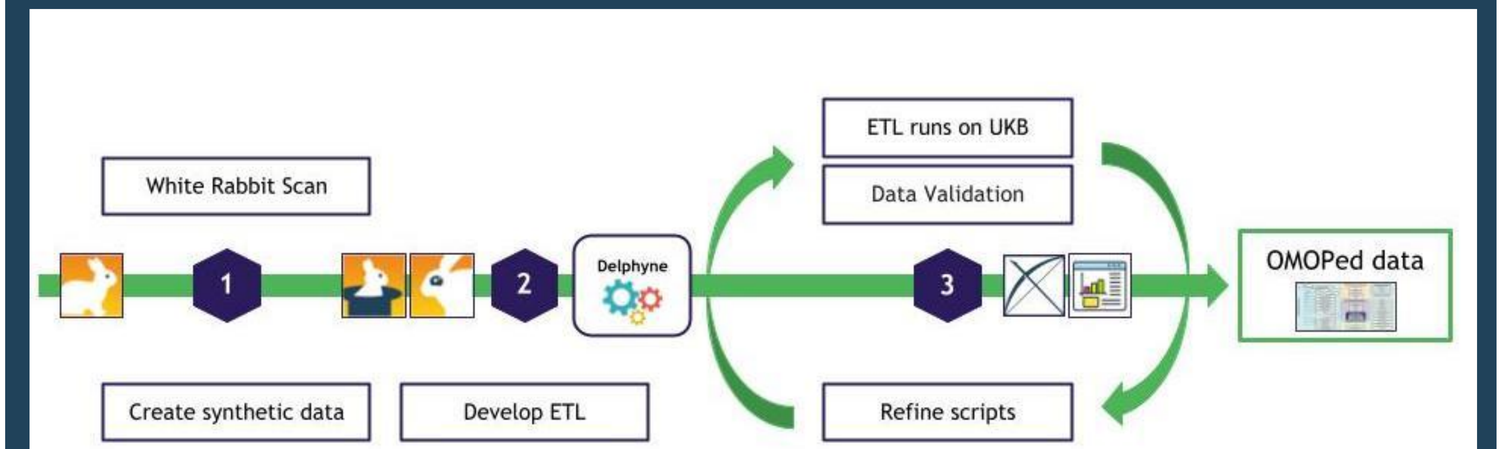

Figure 1. Data conversion workflow using existing OHDSI tools together with our internal ETL pipeline- Delphyne.

Scan QR to **see our blogpost.**

We implemented tests with the R testing framework functionality of RiaH. The code was then deployed at UCL to be executed on the original data. Lastly, we performed quality assessments with both Achilles and the Data Quality Dashboard (DQD).

Results

Our effort to convert the UK Biobank data to the OMOP CDM is an excellent example of successful collaboration and community engagement. Besides working with UCL to enable the UKB data to be used for research, the Hyve actively participated in the OHDSI community, by founding and leading the UKB working group, as well as initiating and taking part in OHDSI Forums discussions on issues that arose from the mapping and would benefit the community as a whole.

In total, the ETL codebase includes 35 table to table transformation scripts, and makes use of 19 OMOP vocabularies for the semantic mapping. The code has been executed successfully at UCL with the DQD currently achieving a 99% pass rate.

Such a high rate is a direct result of regular conversations with UCL, from which we gained useful insights for code adjustments, as well as the use of Delphyne's informative execution logs and summary reports to facilitate the investigation of data quality issues.

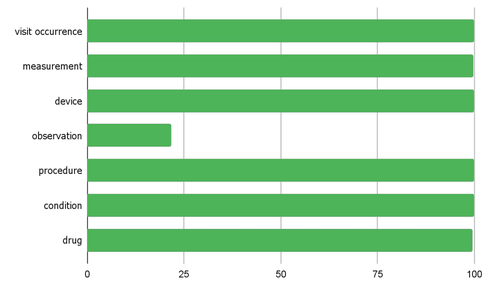

Figure 3. Percentage of UKB source codes mapped to a standard OMOP concept per domain by record frequency based on the full dataset). We achieved a near full or mapping coverage for most tables and the lowest mapping coverage was for the observation.

References

1. UK Biobank. Available from: https://www.ukbiobank.ac.uk/
2. European Health Data Evidence Network (EHDEN). Available from: https://www.ehden.eu/

Sofia Bazakou, Maxim Moinat, Alessia Peviani, Anne van Winzum, Stefan Payralbe, Vaclav Papez, Spiros Denaxas

The Hyve UCL EHDEN EUROPEAN HEALTH DATA & EVIDENCE NETWORK