

Mapping UK Biobank to the OMOP CDM: development of USAGI

PRESENTER: Maxim Moinat

Background

UK Biobank (UKB)¹ is a medical registry containing vast amounts of medical data from 500.000 participants. This registry combines ‘baseline’ data with variables specifically collected for the project with linked primary and hospital EHR data.

Last year, The Hyve collaborated with University College London (UCL) to convert the information into the Observational Medical Outcomes Partnership common data model or OMOP CDM. Here we show how we added and used new features to one of the OHDSI ETL tools: Usagi - utilized in the preparation phase as part of the conversion of the UKB data mapping project. A generalised version of these features was merged back into the OHDSI Usagi Github repository².

The journey to convert data into the OMOP CDM includes various steps. One part of the preparation of the data are the semantic mappings, consisting of translating the (medical) codes in the source data to an existing “standard”, agreed to be used as a common vocabulary. There are different approaches for the semantic mapping, dependent on the available information, and Usagi is used in the cases when the source values are coded in a local vocabulary with no existing mapping to standard concepts available.

In short, Usagi is a Java application that takes all source concepts and finds suitable equivalent “standard” OMOP concepts. The advantage of Usagi is that it searches for similar terms in common vocabularies and classes, and shows a suggested mapping based on a matching score.

In the mapping of the UKB baseline variables, we needed to adapt Usagi to be able to map variable+value pairs and map to value and unit concepts, based on the context of the variable. We will demonstrate our approach here.

We enhanced Usagi to a) map to value and unit target concepts and b) add mapping statuses (flag), c) add mapping provenance and d) mapping equivalence, all making Usagi more suited for mapping the UK Biobank variables.

The screenshot shows the Usagi application interface with several key features highlighted by red boxes and arrows:

- Mapping provenance:** A box highlights the 'Created by' field in the mapping details, showing 'Maxim (2022-03-07)'.
- Mapping types:** A box highlights the 'Mapping type' dropdown menu, which is set to 'MAPS_TO_VALUE'.
- Mapping equivalence:** A box highlights the 'Mapping equivalence' dropdown menu, which is set to 'EQUAL'.
- Mapping status:** A box highlights the 'Flag' dropdown menu, which is set to 'EQUAL'.

The main table displays source terms and target concepts with their respective frequencies and mapping scores. The source term 'Amount of alcohol drunk on a typical drinking day' (20403) is mapped to the target concept 'Number of alcohol units consumed on typical...' (40481011) with a score of 0.59.

Mapping types
 Whilst performing the preparation steps for the UKB mapping, we found that several concepts required pre-coordination between two or more concepts. For example, the field ‘Amount of alcohol drunk on a typical day’ (20403) can take seven values (such as 1=‘1 or 2’ to 5=‘10 or more’). Different from typical OMOP uses cases, the value codes had to be mapped in the context of the variable. This was not supported by Usagi, so we extended Usagi to 1) import *variable+value pairs* and 2) map to *both an event and a value concept*. Following the example, this allowed mapping 20403-2 to concept 40481011 ‘Number of units consumed on a typical drinking day’ and value 4244842 ‘Three to four times a day’.

Mapping provenance
 Usagi can accelerate the semantic mapping step, but it requires a high level of medical expertise. Thus multiple people often need to work with the same file. The Hyve team added metadata fields to Usagi to enhance this collaborative approach. This includes:

- Created by (author)
- Status changed by (author)
- Creation timestamp
- Status change timestamp

Mapping status
 In a similar way, there were mappings that required special attention. For example, mappings for which there is a higher level of uncertainty (‘ill’ could refer to many different conditions), or ones that are not worth retaining (‘patient wears only pink clothes’). We added the ability in Usagi to flag these mappings: a mapping status to signify that the mapping needs further review.

Mapping equivalence
 Moreover, there might not always be an exact mapping to a code. However a higher level term might exist that makes sense for the data. The last addition to the Usagi functionality is to include this ‘level of equivalence’ of mapping. One can choose to label if the mapping is *equal*, *equivalent*, *wider*, *narrower* etc. For example data coding 4 is used in UKB to map treatments. There are two treatments for acne: acnecide 5% gel and acnecide 10% gel and they both map to standard concept benzoyl peroxide. This mapping is to a more general term and this can be made clear in Usagi using ‘WIDER’ for the approved mapping. The equivalence statuses are:

- Equal (identical codes)
- Equivalent (same meaning)
- Wider (target concept contains more information)
- Narrower (target concept contains less information)
- Inexact (target concept contains both more and less information)
- Unreviewed



References

1. UK Biobank. Available from: <https://www.ukbiobank.ac.uk/>
2. OHDSI Usagi. Available from: <https://github.com/ohdsi/Usagi>
3. European Health Data Evidence Network (EHDEN). Available from: <https://www.ehden.eu/>