Conversion of UK Biobank into the OMOP CDM: New Data for Inferences Between Episodic Care

Amelia J Averitt[1]; Alexandra Orlova[2]; Alexander Davydov[2]; Oleg Zhuk[2]; Michael N Cantor[1]; Gregory Klebanov[2]

1. Regeneron Genetics Center. Tarrytown, NY. 2. Odysseus Data Services. Cambridge, MA.

Background. The OMOP Common Data Model (CDM) is central to the mission of OHDSI. The use of CDM-formatted data permits researchers across the OHDSI network to generate meaningfully comparable inferences from the same analyses. Many current CDM-formatted databases are created from electronic health record (EHR) systems or administrative claims data. However, this data is episodic; it is encounter-based and captures little information on the patient state outside of those encounters. Latent patient states between medical encounters may contain important health information. As such, episodic data may be a suboptimal source from which to make inferences for evidence-based care.[2]

This work presents the harmonization and standardization of UK Biobank (UKB) data into the OMOP CDM. The UKB is an important scientific resource that has led to hundreds of publications and discoveries in both genetics and epidemiology.[3,4] It contains data on a large, prospective cohort that includes individuals who were recruited between 2006 and 2010 from 22 assessment centers in England, Scotland, and Wales. Presently, it contains 500,000 individuals who will be followed for a minimum of thirty years. It contains genetic data and in-depth health information including measurements, clinical history, images, survey responses, and limited EHR information.[3,4,5]

The UKB captures a patient trajectory that is normally unobserved. Unlike episodic data, the UKB provides data that speaks to the patient's continual health state. This represents an unprecedented opportunity to create new inferences. By contributing this data to the OHDSI community, we will enable researchers to (i) identify genetic and clinical relationships and (ii) learn about the patient experience between episodes of care. Furthermore, the conversion of the UKB data may serve as a new resource for OHDSI network analyses. Given the demand of this data resource from the scientific community, UKB leadership are supportive of this conversion effort and will make OMOP CDM-formatted UKB data available by request for authorized researchers.

Methods. We sought to convert the non-genetic UKB data into the OMOP CDM v5.3.1. The UKB consists of three datasets - (i) *Main*, (ii) *Primary Care* (PC), and (iii) *Hospital Episode Statistics* (HES).

The Main data captures survey responses of demographic and lifestyle indicators; and baseline measurements of biomarkers of blood, saliva, and urine samples. This data is also linked to cancer and death registries. The PC data captures encounters with general practitioners and is similar to an EHR. Data modalities include diagnoses, history, symptoms, lab results, procedures, and medications. The HES data captures inpatient hospital stays, and provides information on admissions and discharge, diagnoses, surgical procedures,

FIGURE 1. *Demographics of UKB. From left to right: distribution of year of birth, race, and gender*

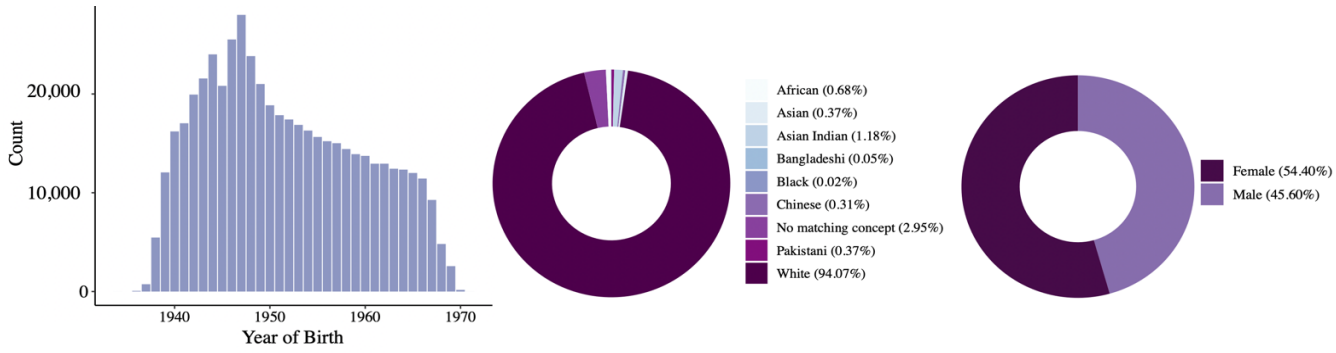

TABLE 1. *Statistics on the ETL procedure to convert UKB data to OMOP CDM.*

OMOP CDM table	total number of records	mapped number of records	mapping rate	unique persons
location	16,062			
provider	152			
care_site	27			
person	502,504			502,504
death	32,831			20,432
observation_period	502,504			502,504
visit_occurrence	3,224,433			502,504
visit_detail	2,568,382			382,524
condition_occurrence	29,523,665	29,410,530	99.62%	502,357
procedure_occurrence	13,710,930	13,682,277	99.79%	442,856
drug_exposure	54,606,521	53,435,779	97.86%	226,481
device_exposure	1,982,280	1,130,005	57.01%	110,475
measurement	172,403,729	34,142,756	19.80%	502,074
observation	734,191,463	422,208,631	57.51%	502,504
specimen	5,453,419	5,453,419	100.00%	497,078
note	17,147			17,147
condition_era	25,198,756	25,198,756	100.00%	502,357
drug_era	19,338,731	19,338,731	100.00%	224,938

*gray areas indicate tables that did not require concept mapping or do not contain individual-level data.

maternity and obstetrics care, and limited psychiatric-related admissions.

These three source datasets were integrated into the single OMOP CDM instance by the process of *extract, transform, load* (ETL). To implement the ETL, *lookup tables* that functionally support logical operations were designed. Logical operations include, but were not limited to (i) the removal of ambiguous records (e.g. a PC record that indicates both the presence and absence of Read code 136P.00 *heavy drinker* was excluded) and (ii) the preservation of content (e.g. a PC record with Read code 246E.00 *Sitting blood pressure* contains both a diastolic and systolic blood pressure measurement, that should each be mapped to OMOP concept codes). A full schematic and greater details on the ETL process can be found at `https://bit.ly/3wJdSTi`

Transforming the survey responses was a unique challenge in this ETL process. Unlike structured clinical data, which can be longitudinally represented by tuples of concept codes and dates, survey responses require transforming data into patient *histories*. Histories were created using either the individual's age or the reported year of the historical event. For example, a survey response of four years for *Time since last prostate specific antigen (PSA)* is mapped to both (i) the occurrence of the measurement - *Past history of procedure* (4215685) and *Prostate specific antigen measurement* (4272032); and (ii) the value of the measurement - *Past history of procedure within 5 years* (907926) and *Past history of procedure longer than 3 years ago* (907959).

Results. Statistics on the ETL can be found in Table 1. A summary of the data can be found in Figure 1. 502,504 unique individuals were mapped into the OMOP CDM. The UKB data was coded using 18 unique vocabularies. The tables `condition_occurrence`, `procedure_occurrence`, and `drug_exposure` had the highest mapping rates (99.62%, 99.79% and 97.86%, respectively), and `measurement` had the lowest mapping rate (19.80%).

Discussion. This conversion highlighted many significant and new data challenges. Notably, the genetic data from the UKB was not converted. Though there are efforts to incorporate genomic data into the OMOP CDM, at present there is no schema to support it.[6] Many advancements to biomedicine, including precision medicine, necessitate the use of genetic data.[7] To fully utilize the potential of this data and produce new biomedical inferences, the OMOP CDM should be able to accommodate this data. We found that the mapping procedure for the PC and HES data was straightforward and typical of ETLs for EHR and administrative claims data. However, the survey responses were more difficult to incorporate into the OMOP CDM, as there is partial information and ambiguity in the responses (e.g. *do not know, do not remember, uncertain*) that cannot be easily accommodated. Additionally, the breadth of the surveys' content covered topics or assessments that have not yet been integrated into standard vocabularies. Usable but non-mappable data points, such as this, contain valuable information that should be retained. We see these issues as opportunities for the OHDSI community to create increasingly flexible and progressive data structures for biomedical research.

In all, the conversion of the non-genetic UKB data to the OMOP CDM is an important contribution to the research community. This effort is a demonstration of the collaborative nature of OHDSI; wherein competitors worked together to provide access to this rich and important data source. This cooperation will enable the broader community to conduct large-scale and reproducible epidemiologic and genetic analyses. Presently, we are working with the UKB to develop a long-term plan for the maintenance of the CDM version as new data is released.

References

1. Rosenbloom, ST et al. Representing Knowledge Consistently Across Health Systems. Yearbook of medical informatics vol. 26,1. 139-147. (2017).
2. Nagykaldi, Zsolt J et al. "Moving From Problem-Oriented to Goal-Directed Health Records." *Annals of family medicine* vol. 16,2. (2018).
3. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. Pharmacogenomics 6: 639–646. (2005).
4. Collins R. What makes UK Biobank special? Lancet 379:1173–1174. (2012).
5. Elliott P, Peakman TC The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int J Epidemiol 37: 234–244. (2008).
6. Genomic CDM Subgroup. Observational Health Data Sciences and Informatics. https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:genetics-sg (2017).
7. Elemento O. The future of precision medicine: towards a more predictive personalized medicine. Emerg Top Life Sci. Sep 8;4(2):175-177. (2020).