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.

Introduction

Many current OMOP 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 sub-optimal source from which to make inferences for evidence-based care. This work presents the harmonization and standardization of UK Biobank (UKB) data into the OMOP CDM.[1][2] Unlike episodic data, the UKB captures a patient trajectory that is normally unobserved and speaks to the patient's continual health state

About the UK Biobank

- A large, prospective cohort of individuals recruited between 2006 and 2010 from 22 assessment centers in England, Scotland, and Wales.
- 500,000+ individuals who will be followed for a minimum of thirty years.
- Consists of genetic data and three clinical datasets.[1][2][3]
  - **Main** - Survey responses for demographics & lifestyle indicators and baseline measurements of blood, saliva, & urine samples. This data is also linked to cancer and death registries.
  - **Primary Care** - Encounters with general practitioners. Data modalities include diagnoses, history, symptoms, lab results, procedures, and medications.
  - **Hospital Episode Statistics** - Inpatient hospital stays, including admissions, discharge, diagnoses, surgical procedures, maternity & obstetrics care, and limited psychiatric-related admissions.

Methods

We sought to convert the non-genetic UKB data into the OMOP CDM v5.3.1. The data was integrated into the single OMOP CDM instance by the process of extract, transform, load (ETL). Some unique aspects of the ETL include -

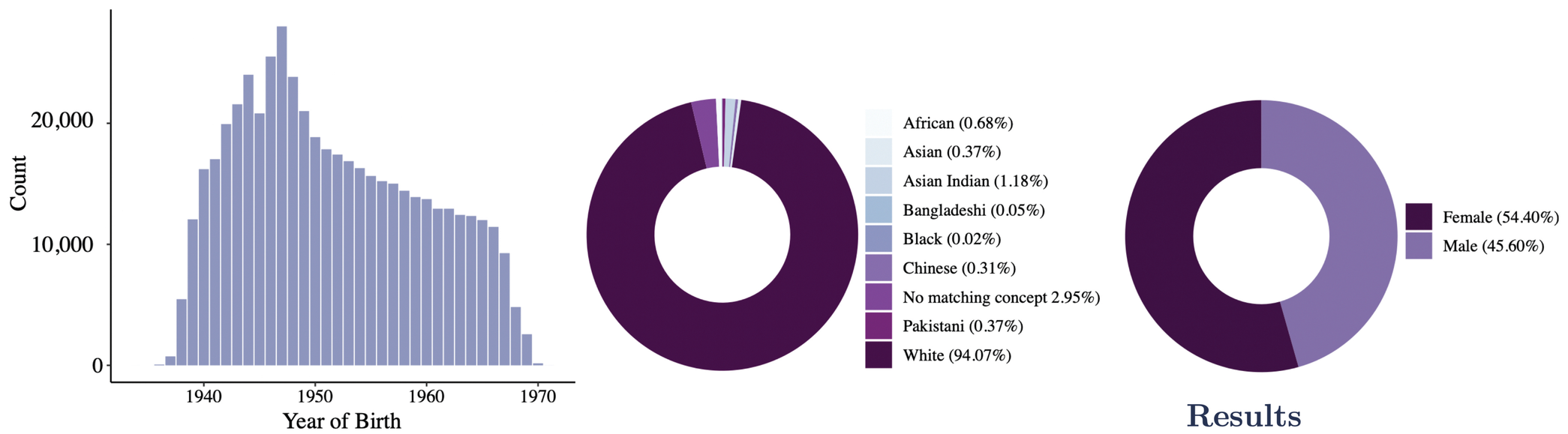
- The use of *lookup tables* to support logical operations.
- The removal of ambiguous records. (E.g. a record that indicates both the presence and absence of Read code 136P.00 heavy drinker was excluded)
- The preservation of content. (E.g. a record with Read code 246E.00 sitting blood pressure contains both a diastolic and systolic blood pressure measurement, that should each be mapped to OMOP concepts).
- The transformation of survey responses into patient *histories*.
  - Patient histories were created using either the individual's age or the reported year of the historical event. (E.g. a survey response of 'four years for time since last prostate specific antigen' is mapped to both (i) the observation of the historical event with the respective event date and (ii) the measurement of the prostate specific antigen as the value of the observation.)

A full schematic and greater details on the ETL process can be found at https://bit.ly/3wJdSTi

Challenges

This conversion highlighted many significant and new data challenges. The genetic data from the UKB was not converted, as there is no clear consensus on representation. The survey responses were difficult to incorporate into the OMOP CDM, as there is partial information and ambiguity in the responses (*do not know*, *do not remember*, *uncertain*) that cannot be easily accommodated.

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



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.

OMOP CDM Table	Total No. of Records	Mapped Np. of Records	Mapping Rate	Unique No. of 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%	502,537
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	34,143,756	19.80%	502,074
observation	734,403,729	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

Table 1. Statistics on the ETL procedure. Gray areas indicate tables that did not require concept mapping or do not contain individual-level data.

Key Take-Aways

- The OMOP CDM-formatted UKB may enable researchers to (i) learn about the patient experience between episodes of care and (ii) explore relationships between genetics and clinical observations.
- This data is available by request for authorized researchers.
- Usable but non-mappable data points contain valuable information that should be retained. This is an opportunity for the OHDSI community to create increasingly flexible and progressive data structures for biomedical research.

[1] Ollier W *UK Biobank: from concept to reality. Pharmacogenomics 6: 639–646. (2005).*

[2] Collins R *What makes UK Biobank special? Lancet 379:1173–1174. (2012).*

[3] Elliott P *The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int J Epidem 37: 234–244. (2008).*

ODYSSEUS DATA SERVICES REGENERON SCIENCE TO MEDICINE biobank uk