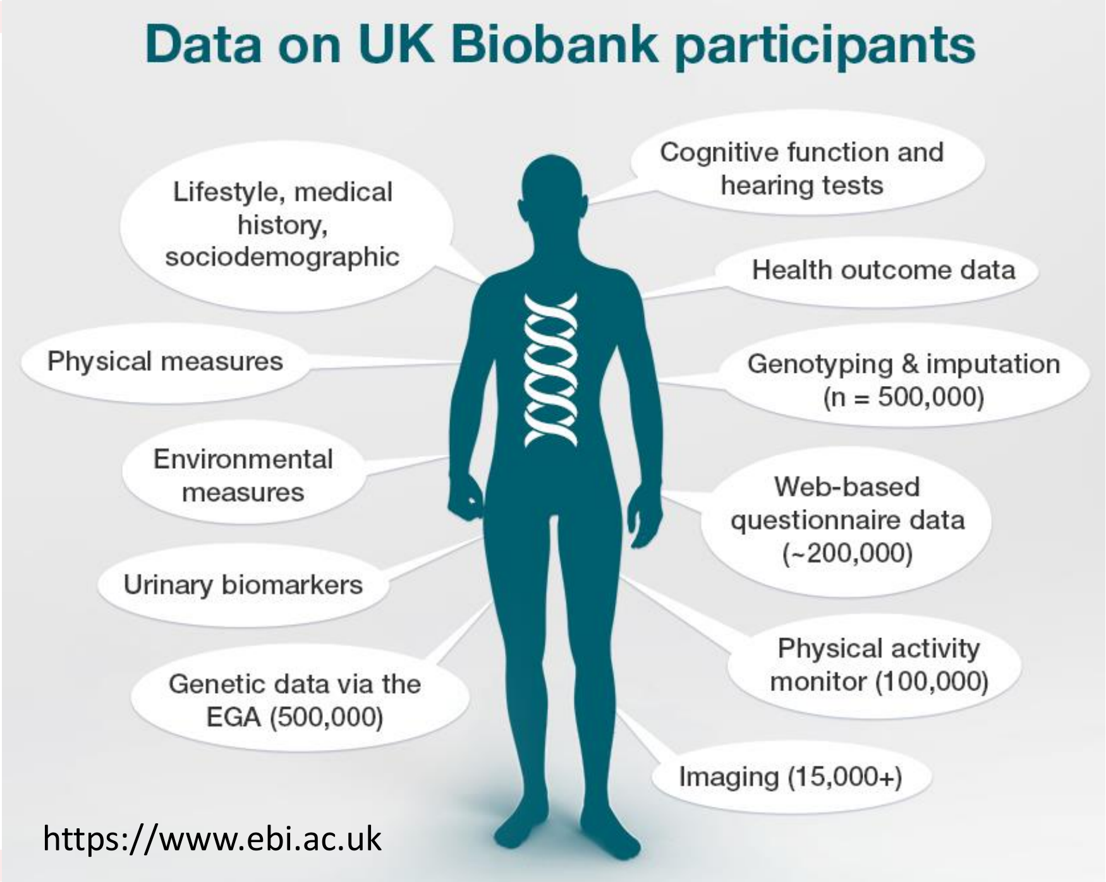
Curation of UK Biobank phenotype data

Emily HM Wong[1], David Ficenec[2], David Merberg[1], Yulia Skovpen[3], Alena Fedarovich[3], Tyler Kolisnik[3], Oksana Tyurina[3], Christine Loreth[3], Mahyar Sabripour[1], Bin Li[1], Yan Ge[1], Tania Khasanova[3], Julie Bryant[3], Sandor Szalma[1]

[1]Data Science Institute and [2]Information Technology, Takeda Pharmaceuticals, USA; [3] Rancho Biosciences, USA

Introduction

- UK Biobank collects vast amounts of phenotype data and provides a valuable resource for studying the complex relationships between the human genome and phenome.
- As part of the UKB-Pharma consortium, we are building a pipeline to process available UK Biobank data. To facilitate downstream analyses, we and Rancho Biosciences first extensively curated phenotype data from all UK Biobank participants.


Data on UK Biobank participants
Lifestyle, medical history, sociodemographic
Cognitive function and hearing tests
Health outcome data
Physical measures
Genotyping & imputation (n = 500,000)
Environmental measures
Web-based questionnaire data (~200,000)
Urinary biomarkers
Physical activity monitor (100,000)
Genetic data via the EGA (500,000)
Imaging (15,000+)
https://www.ebi.ac.uk

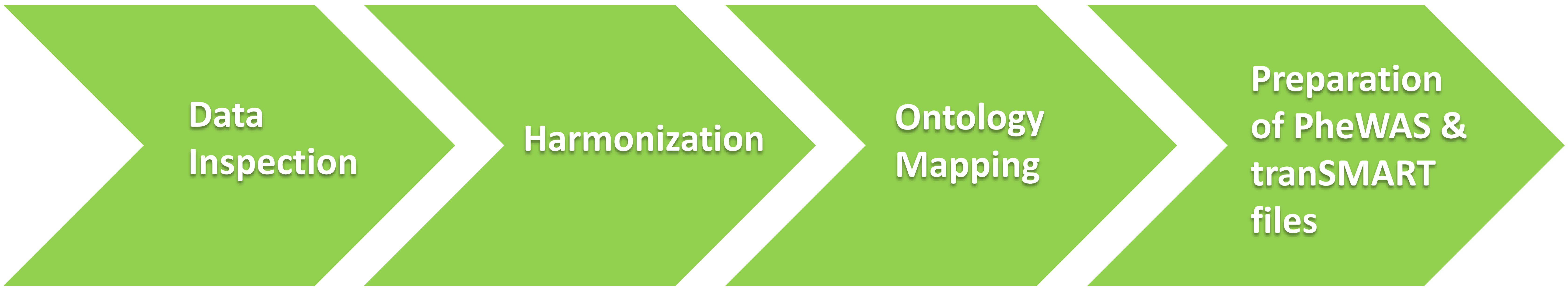
Goal

Curate the UK Biobank data by aggregating and integrating the disparate datasets into a single comprehensive resource

- Flat files (PheWAS ready)
- tranSMART ready files

Curation Process: Overview

Beginning of 2018: **502,616 subjects**
3,390 data fields


Data Inspection → Harmonization → Ontology Mapping → Preparation of PheWAS & tranSMART files

- Data inspection
- UKB coding files update
- Creation of decoded version
- Preparation of supporting files for PheWAS
  - e.g. code reassignment file, field ID–coding file mapping, summary statistics file, etc.

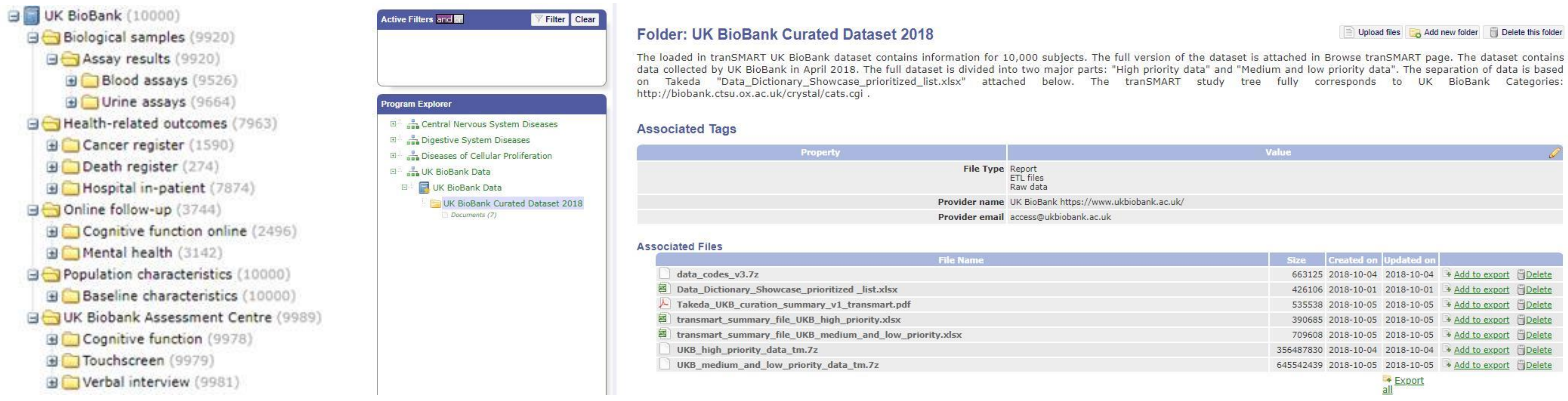
- Harmonization of inconsistent data in multiple instances
- ~100 harmonization rules have been proposed and implemented

- RXNORM
- SNOMED
- ICD10
- MeSH

- Data preparation for PheWAS analysis
- Data conversion into a human-readable format to assist the translational biomedical research
- Upload of UKB phenotype data onto tranSMART

UK Biobank in tranSMART

- The tranSMART platform is a knowledge management architecture for translational medicine.
- It allows for mining and searching multi-model data, including clinical and biomarker data.


- UK BioBank (10000)
  - Biological samples (9920)
    - Assay results (9920)
      - Blood assays (9526)
      - Urine assays (9664)
  - Health-related outcomes (7963)
    - Cancer register (1590)
    - Death register (274)
    - Hospital in-patient (7874)
  - Online follow-up (3744)
    - Cognitive function online (2496)
    - Mental health (3142)
  - Population characteristics (10000)
    - Baseline characteristics (10000)
  - UK Biobank Assessment Centre (9989)
    - Cognitive function (9978)
    - Touchscreen (9979)
    - Verbal interview (9981)

Folder: UK BioBank Curated Dataset 2018

The loaded in tranSMART UK BioBank dataset contains information for 10,000 subjects. The full version of the dataset is attached in Browse tranSMART page. The dataset contains data collected by UK BioBank in April 2018. The full dataset is divided into two major parts: "High priority data" and "Medium and low priority data". The separation of data is based on Takeda "Data_Dictionary_Showcase_prioritized_list.xlsx" attached below. The tranSMART study tree fully corresponds to UK BioBank Categories: http://biobank.ctsu.ox.ac.uk/crystal/cats.cgi .

Associated Tags

Property	Value
File Type	Report ETL files Raw data
Provider name	UK BioBank https://www.ukbiobank.ac.uk/
Provider email	access@ukbiobank.ac.uk

Associated Files

File Name	Size	Created on	Updated on	
data_codes_v3.7z	663125	2018-10-04	2018-10-04	Add to export Delete
Data_Dictionary_Showcase_prioritized _list.xlsx	426106	2018-10-01	2018-10-01	Add to export Delete
Takeda_UKB_curation_summary_v1_transmart.pdf	535538	2018-10-05	2018-10-05	Add to export Delete
transmart_summary_file_UKB_high_priority.xlsx	390685	2018-10-05	2018-10-05	Add to export Delete
transmart_summary_file_UKB_medium_and_low_priority.xlsx	709608	2018-10-05	2018-10-05	Add to export Delete
UKB_high_priority_data_tm.7z	356487830	2018-10-04	2018-10-04	Add to export Delete
UKB_medium_and_low_priority_data_tm.7z	645542439	2018-10-05	2018-10-05	Add to export Delete

Export all

Identifying drug targets and pleiotropy

- Genomics data and the curated phenotype data were loaded onto a Spark and Hail framework on Amazon Databricks.
- This allows us to efficiently perform PHEWAS and GWAS to identify drug targets and pleiotropy in three major therapeutic areas – gastroenterology, neuroscience, and oncology.