Towards a comprehensive view of diagnoses in UK Biobank by data curation and aggregation

Oleg Stroganov[1], Alena Fedarovich[1], Emily Wong[2], Yulia Skovpen[1], Ivan Grishagin[1], Dzmitry Fedarovich[1], Tania Khasanova[1], David Merberg[3], Sándor Szalma[2], Julie Bryant[1]

[1]Rancho BioSciences, LLC, 16955 Via Del Campo #220, San Diego, CA 92127

[2, 3]Takeda Development Center Americas, Inc., 9625 Towne Centre Drive, San Diego, CA 92101 and 35 Lansdowne Street, Cambridge, MA, 02139

The UK Biobank dataset contains phenotypic, genomic, and imaging data on >500,000 participants gathered from questionnaires, bio-sample measurements, assessments, body scans and electronic health records in England, Wales, and Scotland. The dataset continues to grow as the information is continually updated. In September 2019 UK Biobank released new phenotype data including primary care (GP) data for ~ 45% of the cohort, the first occurrence information for a set of diagnoses, and additional brain MRI data. The hospital inpatient (HESIN) data is now available in two formats: summary and record-level. The GP data contains coded prescriptions and clinical information. It brings in new data formats and dictionaries which require significant effort of ontology mapping, reformatting and harmonization for integrating with HESIN and self-reported data. We improved the ontology mapping by making use of publicly available mappings from UK Biobank, TRUD, BioPortal, SNOMED-CT. We also created a curation and data integration pipeline for harmonizing diagnosis, operational procedure and prescription data. It includes automated and manual stages and was used to map >90,000 raw prescription records to MeSH and RxNORM ontologies.

We reviewed several selected diagnoses including ulcerative colitis, Crohn's disease, Alzheimer's disease, bipolar affective disorder, etc. to assess the contribution of each data source (HESIN, self-reported, GP, etc.) used by UK Biobank for patients' diagnoses data.

GP data provides a significant amount (>50% to the number of subjects for 56% of the selected diagnosis codes) of ambulatory care data previously not covered by hospital inpatient data, resulting in a considerable (up to 412%) change in subjects counts for certain diagnoses for example depressive episode, disorders of aromatic amino-acid metabolism, disorders of sphingolipid metabolism and other lipid storage disorders, congenital malformations and deformations, etc.

Analyses revealed issues with automatic mapping of GP data (READ coded) to other ontologies (ICD9, ICD10, OPCS). We observed an excessive number of patients for a number of diagnosis compared to manual mapping due to one-to-many read-to-ICD10 mappings. To overcome these challenges, we leveraged existing TRUD mappings, together with a combination of automated and manual curation.

We conclude that GP data is a valuable source of information for specific diseases. However, when doing phenotypic data analyses researchers may want to exclude subjects with aggregated codes and codes with one-to-many READ-to-ICD10 mappings.