Aggregation and Integration of UK Biobank Phenotype Data for Downstream Analysis

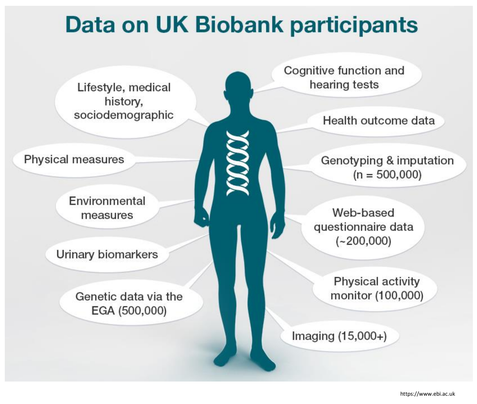
Yulia Skovpen[1], Emily Wong[2], Christine Loreth[1], Alena Fedarovich[1], Tyler Kolisnik[1], Oksana Tyurina[1], Tania Khasanova[1], David Merberg[3], Sándor Szalma[2], Julie Bryant[1]

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

[2, 3]Takeda Pharmaceuticals, 9625 Towne Centre Drive, San Diego, CA 92101 and 35 Lansdowne Street, Cambridge, MA, 02139

Background

The UK Biobank is a rich dataset from 500,000 participants and contains phenotypic data types such as participant questionnaires and HER-derived data, genomic, biomarkers, and imaging data. It is a live dataset that is regularly updated in real-time throughout the year. As a result, data that is not readily amenable for analysis, or consumption accumulates in size. A curation step is required to harmonize incoming data to make it available for downstream data analysis.



- 502,600 subjects
- 3,390 variable fields
- Longitudinal data
- 16,000 data fields

Challenges and Solutions

Inconsistent data at multiple instances	Creation of harmonization rules
Categorical values in “continuous” and “integer” fields	Conversion of categorical values to numerical
“Free text” categorical values; different dictionaries for similar concepts	Application of uniform ontologies
Underpowered cohorts	Algorithmic derivation of new, enriched and scientifically relevant phenotypes

Phenotype Harmonization

Around 100 harmonization rules have been proposed:

- Conversion of years (of events) into subject age, replacement of “ongoing data” by actual values of age or year
- replacement of categorical values in “Continuous” or “Integer” data fields by relevant numeric values
- harmonization of discrepancies in subjects’ answers at multiple visits - when an answer implies a single answer
- define and harmonize data where subject’s answer at the last visit (instance) has most recent information
- identify data fields and subjects with values outside coding files, make appropriate curation decisions

fid	Field name	Data type	Code file	Value range	Code meaning/ Problem	Solution
20420	Longest period spent worried or anxious	Integer	517	-1, 1, 2, 3, …, 839, -999	-999 All my life/as long as I can remember	Replace -999 by Age of subject in months

- **Solution example:** value -999 “All my life/as long as I can remember” to convert into Age of subject at the time of answering the questionnaire

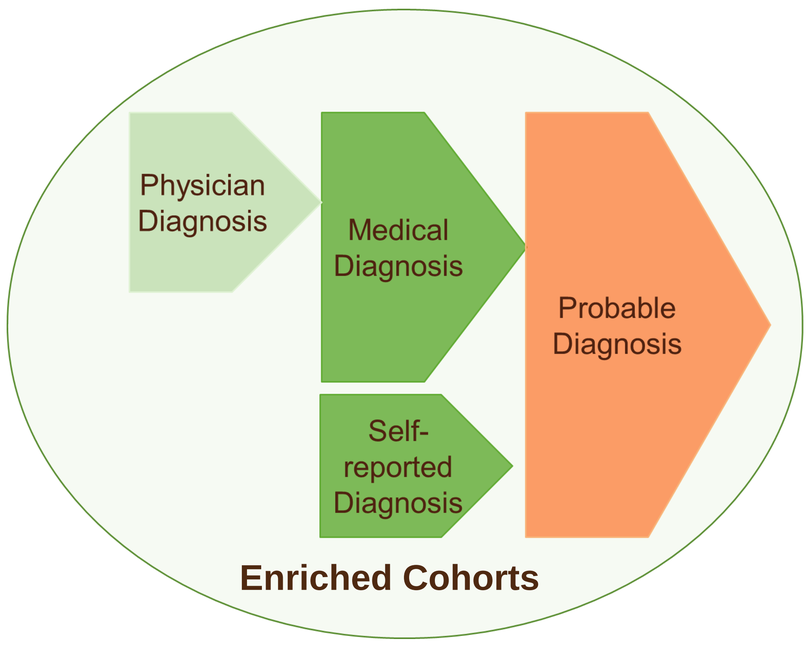
Mapping of Ontologies

- Diagnoses and diseases – ICD10
- Operative procedures – SNOMED CT
- Drugs and medications – RXNORM, MeSH
  - Indication-based drugs classification - MeSH

Category	Count
ear/nose/throat cancer	662
gastrointestinal cancer	3913
neurological system cancer	496
urinary tract cancer	-
kidney/renal cell cancer	713
bladder cancer	1300
other cancer of urinary tract	37
breast cancer	12626
genital tract cancer	11849
haematological malignancy	2822
skin cancer	13512
other cancer	1223
respiratory / intrathoracic cancer	898
unclassifiable	-

ICD10_meaning_coding19	node_id	parent_id	selectable
C03.9 Gum, unspecified	1229	1226	Y
C07 Malignant neoplasm of parotid gland	1247	44	Y
C08.9 Major salivary gland, unspecified	1252	1248	Y
C15.9 Oesophagus, unspecified	1292	1284	Y
C16.9 Stomach, unspecified	1302	1293	Y
C17.9 Small intestine, unspecified	1309	1303	Y
C18.9 Colon, unspecified	1320	1310	Y
C21.0 Anus, unspecified	1324	1323	Y
C18.7 Sigmoid colon	1318	1310	Y
C20 Malignant neoplasm of rectum	1322	45	Y
C22.9 Liver, unspecified	1335	1328	Y
C23 Malignant neoplasm of gallbladder	1336	45	Y
C25.9 Pancreas, unspecified	1350	1342	Y
C34.9 Bronchus or lung, unspecified	1380	1374	Y
C34.9 Bronchus or lung, unspecified	1380	1374	Y
C47.9 Peripheral nerves and autonomic ner	1459	1450	Y
C69.8 Overlapping lesion of eye and adnexa	1561	1553	Y

Enrichment of Cohorts



By combining multiple diagnostic codes from different fields, you can increase the size of disease cohorts. This increase in size will add power to statistics for downstream analysis.

Example of Results

Treatment/medication code mapping to ontologies

ORIGINAL UKB	RXNORM name	RXNORM ID	MESH name	MESH ID
vitamin e product [ctsu]	vitamin e	RXNORM:11256	vitamin e	MESHF:D014810
chondroitin product	chondroitin sulfates	RXNORM:2473	chondroitin	MESHF:D002807
co-enzyme q10/ubiquinone/bio-(	coenzyme q10	RXNORM:21406	coenzyme q10	MESHF:C024989
indigestion remedy (over the cou	unknown/not mapped	NA	unknown/not mapped	NA
omega-3/fish oil supplement	fish oils	RXNORM:4419	fish oils	MESHF:D005395
vitamin c product	ascorbic acid	RXNORM:1151	ascorbic acid	MESHF:D001205
evening primrose oil product	evening primrose oil	RXNORM:203219	efamol	MESHF:C028498
food supplement/plant/herbal e	bioflavonoids	RXNORM:1562	dietary supplements	MESHF:D019587
st john's wort/hypericum [ctsu]	hypericum extract	RXNORM:1309228	hypericum	MESHF:D020902
aloe vera product	aloe extract	RXNORM:91263	aloe	MESHF:D000504
saw palmetto product	serenoa preparation	RXNORM:350488	serenoa	MESHF:D028024
unknown supplement	bioflavonoids	RXNORM:1562	dietary supplements	MESHF:D019587
Free-text entry, unable to be cod	unknown/not mapped	NA	unknown/not mapped	NA
aludrox gel	aluminum hydroxide;magnesium hydrox	RXNORM:612;RXNORM:6581	aluminum hydroxide;mag	MESHF:D000536;MESHF:D008
dijex liquid	aluminum hydroxide;magnesium hydrox	RXNORM:612;RXNORM:6581;RXNO	aluminum hydroxide;mag	MESHF:D000536;MESHF:D008
diovol fruit flavour suspension	aluminum hydroxide;magnesium hydrox	RXNORM:612;RXNORM:6581	aluminum hydroxide;mag	MESHF:D000536;MESHF:D008

Cohorts enrichment

```
f.eid  disease_001  disease_002  disease_003  disease_004  disease_005  disease_006  disease_007  di
5727058  M518  M9596
2339680  M758  M430  I10  E785
3186124  N40  I10  K30  M179  K409  E785
4748085  D061  D069  R91  J304  J189
5046970  C504  Z853  V134  R55  R11  I10  S809  C509
3884101  Z602  J459  Z301  R634  R55  R002  N920  M5446  M512  A099  M518
5429609  Z538  R05  R042  N394  R490  R32  R13  K625  F458  D259
3070554  R31  M8999
4403228  N950  N952
3678822
5374646  J459  G439
1127007  Z864  K660  K562  K529  K219  F412  Z871  R13  Z922  Z720  I10  Z824
```