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]

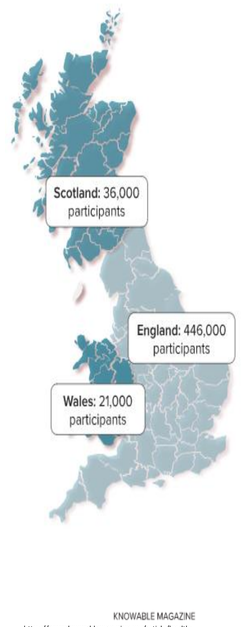
[1]Rancho BioSciences, LLC, 16955 Via Del Campo #220, San Diego, CA 92127

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

OS, AF, YS, IG, DF, TK and JB are Rancho BioSciences affiliated. EW, DM and SS are employees of Takeda.

Background

The UK Biobank dataset contains phenotypic, genomic, and imaging data on >500,000 participants gathered from UK Biobank assessments, questionnaires and electronic health records in England, Wales, Scotland. The dataset continues to grow as the information is continually updated. In September 2019 UK Biobank released more phenotype data including primary care (GP) data for ~ 45% of the cohort, the first occurrence information for a set of diagnoses, and new brain MRI data. The hospital inpatient (HESIN) data has been restructured and updated with a small amount of additional data. The GP category contains coded prescriptions and clinical data and brings in new data formats which require significant effort of data preparation for downstream data analyses.



KNOWABLE MAGAZINE
https://www.knowablemagazine.org/article/health-disease/2019/biobank-research

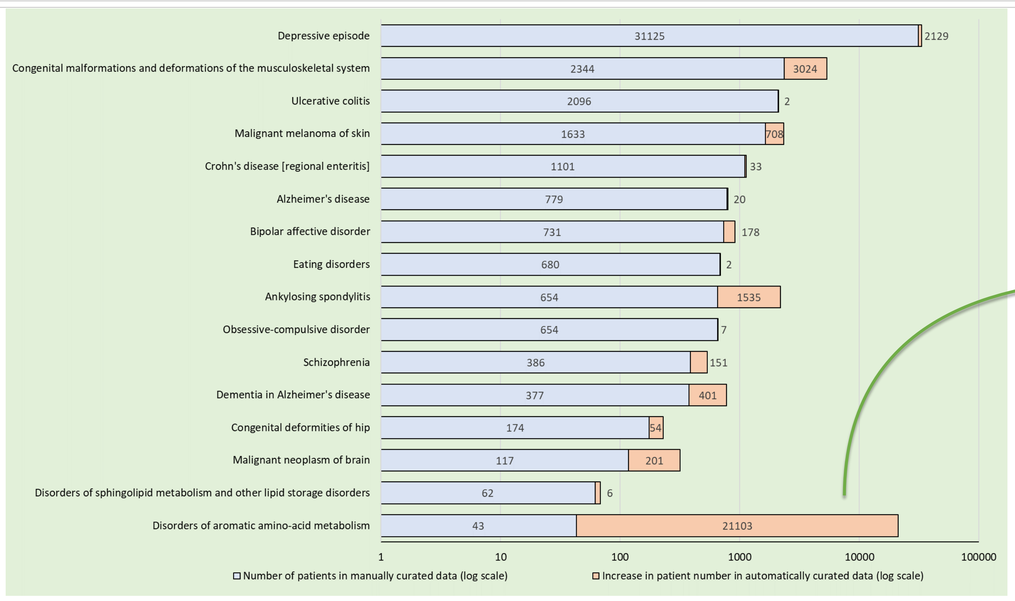
Value added by GP data

GP data contributes significantly (>50%) to the number of subjects for 56% of the selected diagnosis codes. The increase in the number of patients with the selected diagnosis ranged from 7% to 412%.

Disease	ICD10	Full cohort		GP cohort								
		MD	PD	MD	PD	SR	GP	PD or GP	GP & SR and not MD	GP & MD and not SR	GP only	delta, %
Malignant melanoma of skin	C43	4700	6483	2100	2893	1763	1633	3118	178	562	225	7.78
Malignant neoplasm of brain	C71	854	996	341	410	101	117	439	2	74	29	7.07
Crohn's disease [regional enteritis]	K50	2418	2888	1137	1354	811	1101	1597	96	242	243	17.95
Ulcerative colitis	K51	4403	5135	2042	2411	1270	2096	2916	205	619	505	20.95
Disorders of aromatic amino-acid metabolism	E70	28	28	13	13	0	43	47	0	9	34	261.54
Disorders of sphingolipid metabolism and other lipid storage disorders	E75	46	46	19	19	0	62	77	0	4	58	305.26
Alzheimer's disease	F00|G30	1007	1088	441	483	67	779	961	16	265	478	98.96
Dementia in Alzheimer's disease	F00	586	690	282	332	67	377	598	7	96	266	80.12
Schizophrenia	F20	885	1197	362	488	291	386	645	32	91	157	32.17
Bipolar affective disorder	F31	1447	2116	676	1000	671	731	1194	129	151	194	19.40
Depressive episode	F32	18842	65452	8600	30540	26220	31125	47823	8981	2183	17283	56.59
Obsessive-compulsive disorder	F42	226	1248	103	570	494	654	1067	93	44	497	87.19
Eating disorders	F50	137	2171	73	987	951	680	1429	193	14	442	44.78
Ankylosing spondylitis	M45	894	6303	431	2899	2732	654	3108	163	69	209	7.21
Congenital deformities of hip	Q65	148	148	55	55	0	174	210	0	19	155	281.82
Congenital malformations and deformations of the musculoskeletal system	Block Q65-Q79	1172	1173	530	530	0	2344	2718	0	156	2188	412.83

MD and SR were taken from 2019 Q4 data (UKB main dataset). GP diagnosis were taken from gp_clinical primary care dataset, released in 2019
MD - number of unique subjects with medical diagnosis (hesin, cancer registry, death registry)
SR - number of unique subjects with self-reported diagnosis
PD - number of unique subjects with probable diagnosis (MD or SR)
GP - number of unique subjects with diagnosis from GP data (using exact manual curation)
PD or GP - number of unique subjects with probable diagnosis or diagnosis according to GP data
GP & SR and not MD - number of unique subjects who do not have medical diagnosis, and who reported self-diagnosis which was supported by GP data
GP & MD and not SR - number of unique subjects who have medical and GP diagnosis, but who did not report the diagnosis in self-reported data
GP only - number of unique subjects who were diagnosed according to GP data, but do not have corresponding medical or self-reported diagnosis in UKB main dataset
delta, % - increase in number of patients due to addition of GP data to the main dataset [(PD or GP) - PD]/PD, %
Full cohort - all subject in UKB main dataset (502k); GP cohort - patients from gp_clinical primary care dataset (225k)

Challenges in integrating GP data (READ) to diagnoses from other sources (ICD10)



Automated TRUD mapping using selected flags – E (Exact), G (General), DCC or ACC (Default or Alternative mapping, completely refined, no further codes need be added) – gives excessive number of patients for a number of diagnosis compared to manual mapping due to one-to-many read-to-ICD10 mappings.

Automated TRUD mapping has some incorrect mapping between disease and pathology which may lead to excessive number of patients with a diagnosis:

Few read codes that were automatically mapped to "disorders of aromatic amino-acid metabolism" (E70)

Read code	# patients	Description	automated TRUD mapping
XE1DV/READ3	14144	Osteoarthritis	M199;M19;M15;M16;M17;M18;M47;**E702**;M368
N05../READ2	6978	Osteoarthritis and allied disorders	**E702**;M368
other	51		

majority of patients have read3 code XE1DV (Osteoarthritis)

read_ code	icd10_ code	mapping_ status	refine_ flag	add_code_ flag	element_ num	block_ num
XE1DV	M199	D	P	C	0	0
XE1DV	M19	A	M	C	0	0
XE1DV	M15	R	M	C	0	0
XE1DV	M16	R	M	C	0	0
XE1DV	M17	R	M	C	0	0
XE1DV	M18	R	M	C	0	0
XE1DV	M47	R	M	C	0	0
XE1DV	E702D	D	C	C	0	1
XE1DV	M368A	D	C	C	1	1
XE1DV	M368A	D	C	C	0	2
XE1DV	E702D	D	C	C	1	2

Source: Coding system lookups and mapping by UKB/TRUD

Possible solution would be to disregard one-to-many mapping. However, many diagnoses do not have one-to-one mapping and would be lost.

Read codes that were automatically or manually mapped to "ankylosing spondylitis" (M45)

Read code	# patients	Description	automated TRUD mapping
N0401/READ3	1	Other rheumatoid arthritis of spine	M45;M081
N1460/READ3	1	Lumbosacral ankylosis	M432;M45
N1461/READ3	2	Sacroiliac ankylosis	M432;M45
XE1DD/READ3	2185	Musculoskeletal disorder	M00-M90
N100./READ2	190	Ankylosing spondylitis	J998A
N100./READ3	463	Ankylosing spondylitis	J998A

precise code is N100. (Ankylosing spondylitis)

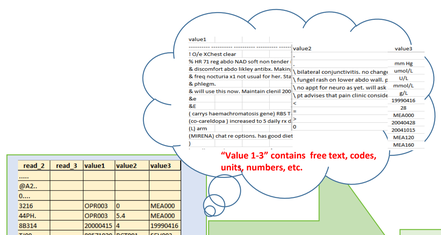
Precise mapping is missing in automatic TRUD mapping because other term (J998A, "Respiratory disorders in other diseases classified elsewhere") have "completely refined" flag and is therefore preferred. Only one-to-many mappings exist in TRUD.

read_ code	icd10_ code	mapping_ status	refine_ flag	add_code_ flag	element_ num	block_ num
N100.	M45X	D	P	C	0	0
N100.	M45XD	D	P	C	0	1
N100.	J998A	D	C	C	1	1
N100.	J998A	D	C	C	0	2
N100.	M45XD	D	P	C	1	2

Manual curation is required for precise READ-ICD10 and READ-OPCS4 mappings

A. GP clinical before and after curation (example)

- 230,105 participants
- All clinical events READ2/3 coded
- Longitudinal data (123,7M records)



read_2	read_3	value1	value2	value3
.....				
@A2..				
0.....				
3216		OPR003	0	MEA000
44PH.		OPR003	5.4	MEA000
8B314		20000415	4	19990416
TJ00.		80571020	RCT001	SEV003
TJz..		ZZZZZZZZ	RCT001	SEV003
Az....				
F12z.				
70001				
70002				
	F12z.			
	F11y1			
	F121.			
	F11z.			
	F12..			
	F121.			
	F223.			
	F23..			
	F340.			
B89E0				
ZE3..				
ZE87.				
7110.				
7K2..				

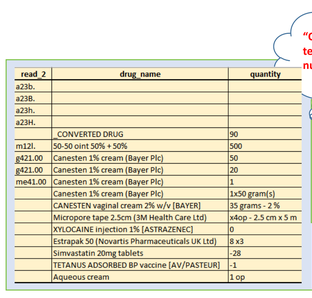
1. READ to Description, ICD9, ICD10, OPCS4:
- UK Biobank mappings
- UK NHSD TRUD mappings

2. ICD9 to ICD10; OPCS4 to SNOMED
3. Unmapped READs
- categorize description
- map to ICD10 or SNOMED

read_2	read_3	Description TRUD	ICD-10 code	SNOMED	Disease/Operation mapping source
.....					
@A2..					
0.....		Occupations			
3216		ECG normal			
44PH.		Total cholesterol measurement			
8B314		Medication review			
TJ00.		Adverse reaction to penicillins	Y400		all_lkps_maps.xlsx (UK Biobank)
TJz..		Adverse reaction to drug NOS	Block Y40-Y59		all_lkps_maps.xlsx (UK Biobank)
Az....		Infectious and parasitic diseases NOS	A00-B99		all_lkps_maps.xlsx (UK Biobank)
F12z.		Parkinson's disease NOS	G20		all_lkps_maps.xlsx (UK Biobank)
70001		Total lobectomy of brain		171443003	all_lkps_maps.xlsx (UK Biobank)
70002		Partial lobectomy of brain		67402009	all_lkps_maps.xlsx (UK Biobank)
	F12z.	Parkinson's disease NOS	G20		TRUD icd9, icd10 mapping
	F11y1	Cerebral ataxia	G119		TRUD icd9, icd10 mapping
	F121.	Drug-induced parkinsonism	G211;G210		TRUD icd9, icd10 mapping
	F11z.	Cerebral degeneration: [NOS]++[Cerebral degeneration: [NOS] or [cerebral atrophy]	G319		TRUD icd9, icd10 mapping
	F12..	Parkinson's disease	G20;G21;G22		TRUD icd9, icd10 mapping
	F121.	Drug-induced parkinsonism	G211;G210		TRUD icd9, icd10 mapping
	F223.	Right: [hemiplegia] or [side++]Right: [hemiplegia] or [sided weakness]	G81		mapped manually
	F23..	Congenital cerebral palsy (&++]Congenital cerebral palsy (& spastic)	G809		mapped manually
	F340.	[Carpal tunnel syndrome) or ++] [Carpal tunnel syndrome) or [partial thenar atrophy]	G56		mapped manually
B89E0		[M]Intraductal carcinoma and lobular carcinoma in situ	D050;D051		mapped manually
ZE3..		Distorted hearing	H93		mapped manually
ZE87.		Hearing loss	H919		mapped manually
7110.		Thyroidectomy operations		13619001	mapped manually
7K2..		Hip joint operations		28181007	mapped manually

B. GP prescriptions before and after curation (example)

- 222,122 participants
- 90,000+ uncleaned prescription records
- Longitudinal data (57,7M records)



read_2	drug_name	quantity
a23b.		
a23B.		
a23h.		
a23H.		
	_CONVERTED DRUG	90
m12l.	50-50 oint 50% + 50%	500
g421.00	Canesten 1% cream (Bayer Plc)	50
g421.00	Canesten 1% cream (Bayer Plc)	20
me41.00	Canesten 1% cream (Bayer Plc)	1
	Canesten 1% cream (Bayer Plc)	1x50 gram(s)
	CANESTEN vaginal cream 2% w/v [BAYER]	35 grams - 2 %
	Micropore tape 2.5cm (3M Health Care Ltd)	x4op - 2.5 cm x 5 m
	XYLOCAINE injection 1% [ASTRAZENEC]	0
	Estrapak 50 (Novartis Pharmaceuticals UK Ltd)	8 x3
	Simvastatin 20mg tablets	-28
	TETANUS ADSORBED BP vaccine [AV/PASTEUR]	-1
	Aqueous cream	1 op

1. READ2 to drug name:
- UK Biobank mapping

2. Map drug name to MeSH/RxNORM

read_2	Drug name	DRUG_CODE	MESH Name	MESH ID	RXNORM Name	RXNORM ID
a23b.	GASTROCOTE tablets	DRUG10707	alginic acid;aluminium hydroxide;magnesium trisilicate;sodium bicarbonate	MESHF:C008970;MESHF:D000536; MESHF:C001671;MESHF:D017693	alginic acid;aluminum hydroxide;magnesium trisilicate;sodium bicarbonate	RXNORM:17305;RXNORM:612;RXNORM:29170;RXNORM:36676
a23B.	GAVISCON peppermint liquid 150ml	DRUG10707	alginate;aluminium hydroxide;magnesium trisilicate;sodium bicarbonate drug combination	MESHF:C007829	alginic acid;aluminum hydroxide;magnesium trisilicate;sodium bicarbonate	RXNORM:17305;RXNORM:612;RXNORM:29170;RXNORM:36676
a23h.	INFACOL liquid	DRUG10017	simethicone	MESHF:D012841	simethicone	RXNORM:9796
a23H.	GAVISCON COOL tablets	DRUG10707	alginate;aluminium hydroxide;magnesium trisilicate;sodium bicarbonate drug combination	MESHF:C007829	alginic acid;aluminum hydroxide;magnesium trisilicate;sodium bicarbonate	RXNORM:17305;RXNORM:612;RXNORM:29170;RXNORM:36676
	_CONVERTED DRUG	DRUG10003	unknown/not mapped	NA	unknown/not mapped	NA
m12l.	50-50 oint 50% + 50%	DRUG10003	dermatologic agents	MESHF:D003879	unknown/not mapped	NA
g421.00	Canesten 1% cream (Bayer Plc)	DRUG10915	clotrimazole	MESHF:D003022	clotrimazole	RXNORM:2623
me41.00	Canesten 1% cream (Bayer Plc)	DRUG10915	clotrimazole	MESHF:D003022	clotrimazole	RXNORM:2623
	Canesten 1% cream (Bayer Plc)	DRUG10915	clotrimazole	MESHF:D003022	clotrimazole	RXNORM:2623
	CANESTEN vaginal cream 2% w/v	DRUG10915	clotrimazole	MESHF:D003022	clotrimazole	RXNORM:2623
	Micropore tape 2.5cm (3M Health	OTHER01	equipment and supplies	MESHF:D004864	unknown/not mapped	NA
	XYLOCAINE injection 1%	DRUG11690	lidocaine	MESHF:D008012	lidocaine	RXNORM:6387
	Estrapak 50 (Novartis	DRUG10657	estradiol;norethindrone	MESHF:D004958;MESHF:D009640	estradiol;norethindrone	RXNORM:4083;RXNORM:7514
	Simvastatin 20mg tablets	DRUG10574	simvastatin	MESHF:D019821	simvastatin	RXNORM:36567
	TETANUS ADSORBED BP vaccine [AV/PASTEUR]	DRUG12789	tetanus toxoid	MESHF:D013745	tetanus toxoid vaccine, inactivated	RXNORM:798306
	Aqueous cream	DRUG11309	emulsifying agents;mineral oil;paraffin	MESHF:D045703;MESHF:D008899; MESHF:D010232	mineral oil;paraffin	RXNORM:6972;RXNORM:7906

C. GP clinical disease and operations mapping challenges and results

Curation challenges

UK Biobank mappings for ICD9, ICD10, OPCS4 contain distorted formats

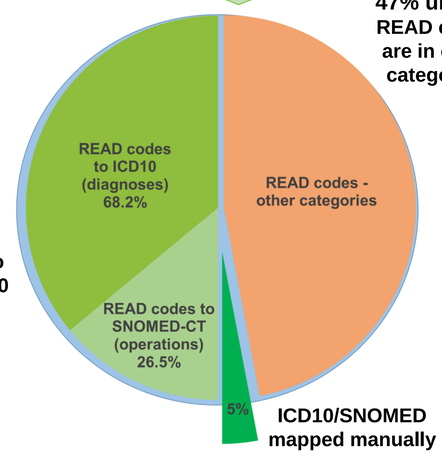
Broad disease and operation codes, ICD9/10 in a different format

READ codes have one-to-many decoding

Both UK Biobank and TRUD READ to ICD9, ICD10, OPCS4 mappings are not complete

read_code	icd10_code	icd9_code	opcs4	Description	ICD10	SNOMED
1.30E+42	R948					
1.30E+51	R948					
1.30E+72	R942					
A77..		77		Other diseases of conjunctiva due to viruses	077	
A77X.		779		Viral conjunctivitis, unspecified	0779	
A771.	B300D H192A	0771D 3704A		Epidemic keratoconjunctivitis	B300;H192	
M22..	L80-L89	702		Other dermatoses	L80-L89	
A.....		001-139		Infectious and parasitic diseases	A00-B99	
AD2..		130-136		Other infectious or parasitic disease NOS	Block B99-B99;Block B50-B64	
70224			A279+G403	Vagotomy & pyloroplasty		65310005;363742001
784z.			M72-M83	Urethra and other urinary operations NOS		15197002;392064002
7A696			L878/Y114	Endovascular radiofrequency ablation of varicose vein of leg		274023003;702590004

Leverage existing TRUD mappings, together with a combination of automated and manual curation



47% unique READ codes are in other categories

53% unique READ codes were mapped to diagnoses ICD10 and operations SNOMED-CT

ICD10/SNOMED mapped manually

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.