

Automated extractions of some phenotypic data tables from the UKBiobank

Futurologist

February 24, 2020

Introduction

- ▶ **UKBiobank:** Genotypic data, Phenotypic data (such as health condition history, imaging, blood tests, etc) and Environmental factors of nearly 500000 people;
- ▶ Possibly to reduce the file-sizes, the raw phenotypic data in the UKB is in a format not suitable for statistical analysis
- ▶ It needs to be converted to more useful data tables, e.g. tables with binary entries for binary categorical variables.
- ▶ I have tried to write some R code that converts some subtables of the raw UKB data-files to formats more suitable for statistical analysis.

UK Biobank Documentation Website: Homepage

The screenshot shows the UK Biobank Documentation Website homepage. At the top, a search bar contains the URL "biobank.ctsu.ox.ac.uk/showcase/index.cgi". Below the search bar is a navigation bar with links to "Index", "Browse", "Search", "Catalogues", "Downloads", "Login", and "Help". The main content area features a welcome message and a list of links with descriptions:

- Essential Information**: Information regarding timelines, updates, release schedules etc.
- Browse**: Find data items by navigating according to their category of origin.
- Search**: Find data items by searching on keywords and other characteristics.
- Catalogues**: Simple listings of database contents and additional resources.
- Downloads**: Download supporting utilities.
- Login**: Apply for access and enable data download.

At the bottom, a legal notice states: "Without a written licence from UK Biobank, you may not copy, reproduce, republish, download, distribute, make available to the public or otherwise use any of the content displayed on this website in whole or in part or permit or assist any third party to do the same, except to the extent permitted at law." The footer also includes the slogan "Improving the health of future generations" and a set of navigation icons.

biobank^{uk}

Welcome to the online showcase of UK Biobank resources. If you are new to using the showcase we recommend you begin by reading the short introductory **User Guide**. Please note that the showcase contains only anonymous summary information.

- Essential Information**
Information regarding timelines, updates, release schedules etc.
- Browse**
Find data items by navigating according to their category of origin.
- Search**
Find data items by searching on keywords and other characteristics.
- Catalogues**
Simple listings of database contents and additional resources.
- Downloads**
Download supporting utilities.
- Login**
Apply for access and enable data download.

Legal notice: Without a written licence from UK Biobank, you may not copy, reproduce, republish, download, distribute, make available to the public or otherwise use any of the content displayed on this website in whole or in part or permit or assist any third party to do the same, except to the extent permitted at law.

Improving the health of future generations

UK Biobank Structure: Top Level

At a **UKB Assessment Center**, the patients filled out a **touchscreen** questionnaire and then some results were discussed with a health-care professional during a **verbal interview**.

biobank ^{uk}

Index Browse Search Catalogues Downloads Login Help

Browse by Primary Category of Origin

Category	Items
Population characteristics	35
UK Biobank Assessment Centre	0
Recruitment	20
Touchscreen	396
Verbal interview	0
Early life factors	5
Employment	3
Medical conditions	13
Medications	5
Operations	7
Physical measures	413
Cognitive function	102
Imaging	2403
Biological sampling	10
Procedural metrics	74
Biological samples	466
Genomics	94
Online follow-up	820
Additional exposures	227
Health-related outcomes	2470

Top Level
Level 1
Level 2
Level 3

Summary generated 10 February 2020

Improving the health of future generations

UK Biobank Assessment Center → Verbal Interview → Medical Conditions

Description

This category contains data obtained through a verbal interview by a trained nurse on past and current medical conditions, including type of cancer and other illnesses, the number of medical conditions, and date of diagnosis.

The interviewer was made aware via a pop-up box on their computer screen if the participant had answered in the touchscreen that they had a history of one or more of the following illnesses: heart attack, angina, stroke, high blood pressure, blood clot in leg, blood clot in lung, emphysema/chronic bronchitis, asthma or diabetes, and was prompted to confirm these with the participant (these will already be selected in the illness screen if they had been selected during the touchscreen questionnaire). If during the interview it appeared these had been incorrectly selected, the interviewer could amend the responses. If the participant stated in the touchscreen they had no major illnesses or disability or were not sure, this question was asked again and confirmed by the interviewer.

Medical conditions that could not be assigned a code at the time of the interview were entered as free text, and subsequently coded wherever possible.

13 Data-Fields

1 Parent Category

3 Resources

Field ID Description

20001	Cancer code, self-reported
84	Cancer year/age first occurred
20007	Interpolated Age of participant when cancer first diagnosed
20009	Interpolated Age of participant when non-cancer illness first diagnosed
20006	Interpolated Year when cancer first diagnosed
20008	Interpolated Year when non-cancer illness first diagnosed
20012	Method of recording time when cancer first diagnosed
20013	Method of recording time when non-cancer illness first diagnosed
20002	Non-cancer illness code, self-reported
87	Non-cancer illness year/age first occurred
134	Number of self-reported cancers
135	Number of self-reported non-cancer illnesses
3140	Pregnant

Data-Field 20002: Self-Reported Medical Conditions

Check: Data, Related Data-Fields, instancing 2, **Data-Coding 6**, Notes

Data-Field 20002

Description: Non-cancer illness code, self-reported

Category: [Medical conditions - Verbal interview - UK Biobank Assessment Centre](#)

Participants	385,697
Item count	1,121,528
Stability	Complete

Value Type	Categorical (multiple)
Item Type	Data
Strata	Derived

Sexed	Both sexes
Instances	Defined (4)
Array	Yes (34)

Debut	Jan 2012
Version	Feb 2020


Data **Notes** **5 Categories** **5 Related Data-Fields** **0 Tabulations** **3 Resources**

1,121,528 items of data are available, covering 385,697 participants, encoded using Data-Coding 6.
Defined-instances run from 0 to 3, labelled using Instancing 2.
Array indices run from 0 to 33.

Category	Count	
cardiovascular	303728	Top level
respiratory/ent	121414	
gastrointestinal/abdominal	114200	
renal/urology	31692	
endocrine/diabetes	65796	
neurology/eye/psychiatry	123602	Level 1
musculoskeletal/trauma	164402	Level 2
haematology/dermatology	35084	Level 3
gynaecology/breast	37322	
immunological/systemic disorders	55890	Level 4
infections	38328	
unclassifiable	-	

Counts of participants/items last updated 10 Jan 2020.

Data-Field 20002: Related Data-Fields



[Index](#)[Browse](#)[Search](#)[Catalogues](#)[Downloads](#)[Login](#)[Help](#)

Data-Field 20002

Description: Non-cancer illness code, self-reported

Category: [Medical conditions - Verbal interview - UK Biobank Assessment Centre](#)


Participants	385,697	Value Type	Categorical (multiple)	Sexed	Both sexes	Debut	Jan 2012
Item count	1,121,528	Item Type	Data	Instances	Defined (4)	Version	Feb 2020
Stability	Complete	Strata	Derived	Array	Yes (34)		

[Data](#)[Notes](#)[5 Categories](#)[5 Related Data-Fields](#)[0 Tabulations](#)[3 Resources](#)

Field ID	Description	Relationship
20009	Interpolated Age of participant when ...	Field 20009 indicates the participant age for the non-cancer illness in Current Field
20008	Interpolated Year when non-cancer il ...	Field 20008 indicates the year of the non-cancer illness in Current Field
20013	Method of recording time when non-ca ...	Field 20013 indicates if date or age was recorded for Current Field
87	Non-cancer illness year/age first oc ...	Field 87 indicates the occasion of the non-cancer illness in Current Field
135	Number of self-reported non-cancer i ...	Field 135 is a count of the number of items for Current Field

Improving the health of future generations

Data-Field 20002: Instancing (number of visits)



[Index](#)[Browse](#)[Search](#)[Catalogues](#)[Downloads](#)[Login](#)[Help](#)

Instance 2

Name: Assessment Centre Visit

All participants attended an initial assessment centre. A proportion were invited several years later to repeat the assessment. For some participants the repeat visit also captured information which was not gathered during their initial visit.

4 Instances

3939 Data-Fields

Index	Description
0	Initial assessment visit (2006-2010) at which participants were recruited and consent given
1	First repeat assessment visit (2012-13)
2	Imaging visit (2014+)
3	First repeat imaging visit (2019+)

Improving the health of future generations

Data-Field 20002

Data Notes 5 Categories 5 Related Data-Fields 0 Tabulations 3 Resources

1,121,528 items of data are available, covering 385,697 participants, encoded using Data-Coding 6.
Defined-instances run from 0 to 3, labelled using Instancing 2.
Array indices run from 0 to 33.

Category	Count	
cardiovascular	303728	
respiratory/ent	-	Top level
asthma	65125	
chronic obstructive airways disease/copd	2075	Level 1
emphysema/chronic bronchitis	7300	
bronchitis	4449	Level 2
emphysema	217	
alpha-1 antitrypsin deficiency	25	
bronchiectasis	1374	Level 3
interstitial lung disease	624	
other respiratory problems	6409	Level 4
ent disorder/not cancer	25256	
respiratory infection	8560	
gastrointestinal/abdominal	114200	
renal/urology	31692	
endocrine/diabetes	65796	
neurology/eye/psychiatry	123602	
musculoskeletal/trauma	164402	
haematology/dermatology	35084	
gynaecology/breast	37322	
immunological/systemic disorders	55890	
infections	38328	
unclassifiable	-	

Data-Field 20002: Data-Codes 6

Data-Coding 6

Name: Non-cancer illness

Description: Tree-structured list used by clinic nurses to code non-cancer illness

Note that myasthenia gravis appears twice (under codes 1260 and 1437) and both these should be treated identically.

This is a hierarchical tree-structured dictionary which uses integers to represent categories or special values.

Coding can be downloaded here as a tab-separated file.

[Download](#)

474 Categories 1 Data-Field

Some items were NOT selectable

Coding	Meaning	Selectable	Node	Parent
-1 cardiovascular		No	1071	Top
-1 respiratory/ent		No	1072	Top
-1 gastrointestinal/abdominal		No	1073	Top
-1 renal/urology		No	1074	Top
-1 endocrine/diabetes		No	1075	Top
-1 neurology/eye/psychiatry		No	1076	Top
-1 musculoskeletal/trauma		No	1077	Top
-1 haematology/dermatology		No	1078	Top
-1 gynaecology/breast		No	1079	Top
-1 immunological/systemic disorders		No	1080	Top
1065 hypertension		Yes	1081	1071
1066 heart/cardiac problem		Yes	1082	1071
-1 cerebrovascular disease		No	1083	1071
1067 peripheral vascular disease		Yes	1084	1071
1068 venous thromboembolic disease		Yes	1085	1071
1072 essential hypertension		Yes	1089	1081

1072



Highlight all

Match case

Match Diacritics

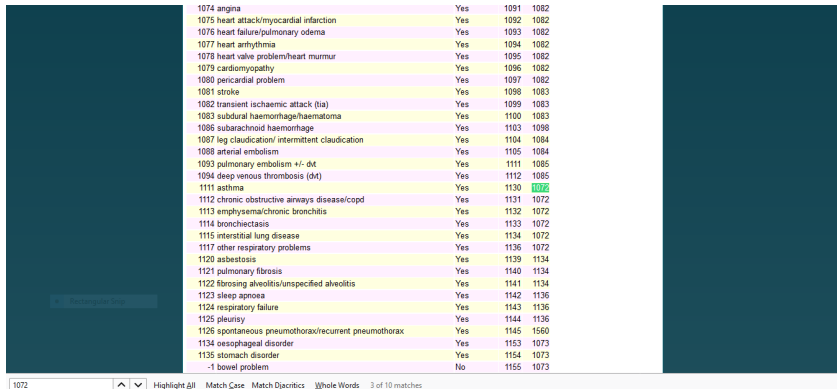
Whole Words

1 of 10 matches

Data-Field 20002: Data-Codes 6

Data-Field Code: header of a UKB data table (e.g. 20002)

Data-Code: entry in a UKB data table (e.g. 1111 for Asthma)



1074 angina	Yes	1091	1082
1075 heart attack/myocardial infarction	Yes	1092	1082
1076 heart failure/pulmonary odema	Yes	1093	1082
1077 heart arrhythmia	Yes	1094	1082
1078 heart valve problem/heart murmur	Yes	1095	1082
1079 cardiomyopathy	Yes	1096	1082
1080 pericardial problem	Yes	1097	1082
1081 stroke	Yes	1098	1083
1082 transient ischaemic attack (tia)	Yes	1099	1083
1083 subdural haemorrhage/haematoma	Yes	1100	1083
1086 subarachnoid haemorrhage	Yes	1103	1098
1087 leg claudication/ intermittent claudication	Yes	1104	1084
1088 arterial embolism	Yes	1105	1084
1093 pulmonary embolism +/- dvt	Yes	1111	1085
1094 deep venous thrombosis (dvt)	Yes	1112	1085
1111 asthma	Yes	1130	1072
1112 chronic obstructive airways disease/copd	Yes	1131	1072
1113 emphysema/chronic bronchitis	Yes	1132	1072
1114 bronchiectasis	Yes	1133	1072
1115 interstitial lung disease	Yes	1134	1072
1117 other respiratory problems	Yes	1136	1072
1120 asbestosis	Yes	1139	1134
1121 pulmonary fibrosis	Yes	1140	1134
1122 fibrosing alveolitis/unspecified alveolitis	Yes	1141	1134
1123 sleep apnoea	Yes	1142	1136
1124 respiratory failure	Yes	1143	1136
1125 pleurisy	Yes	1144	1136
1126 spontaneous pneumothorax/recurrent pneumothorax	Yes	1145	1560
1134 oesophageal disorder	Yes	1153	1073
1135 stomach disorder	Yes	1154	1073
-t bowel problem	No	1155	1073

1072 ^ Highlight all Match Case Match Diacritics Whole Words 3 of 10 matches

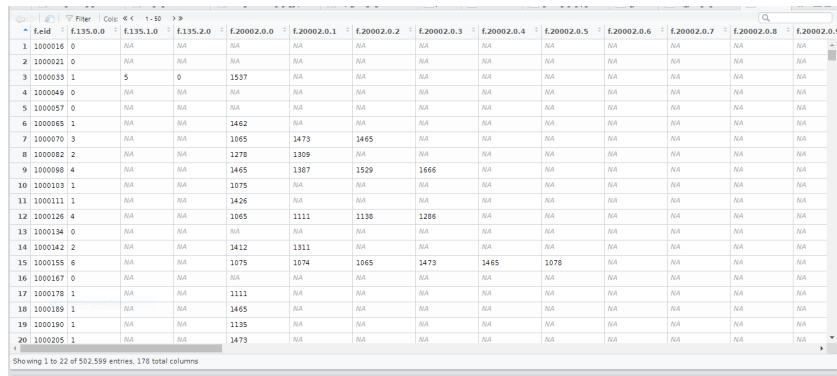
A Raw UKB Subtable: Example

Data-Field Code: header of a UKB data table (e.g. 20002)

Data-Code: entry in a UKB data table (e.g. 1111 for Asthma)

Here: Related Data-Fields 135 (number of self-reported conditions)

20002 (self-reported conditions) and 20009 (age of onset)



	f.field	f.135.0.0	f.135.1.0	f.135.2.0	f.20002.0.0	f.20002.0.1	f.20002.0.2	f.20002.0.3	f.20002.0.4	f.20002.0.5	f.20002.0.6	f.20002.0.7	f.20002.0.8	f.20002.0.9
1	1000016	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	1000021	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	1000033	1	5	0	1537	NA	NA	NA	NA	NA	NA	NA	NA	NA
4	1000049	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	1000057	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6	1000065	1	NA	NA	1462	NA	NA	NA	NA	NA	NA	NA	NA	NA
7	1000070	3	NA	NA	1065	1473	1465	NA	NA	NA	NA	NA	NA	NA
8	1000082	2	NA	NA	1278	1309	NA	NA	NA	NA	NA	NA	NA	NA
9	1000098	4	NA	NA	1465	1387	1529	1666	NA	NA	NA	NA	NA	NA
10	1000103	1	NA	NA	1075	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	1000111	1	NA	NA	1426	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	1000126	4	NA	NA	1065	1111	1138	1286	NA	NA	NA	NA	NA	NA
13	1000134	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	1000142	2	NA	NA	1412	1311	NA	NA	NA	NA	NA	NA	NA	NA
15	1000155	6	NA	NA	1075	1074	1065	1473	1465	1078	NA	NA	NA	NA
16	1000167	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
17	1000178	1	NA	NA	1111	NA	NA	NA	NA	NA	NA	NA	NA	NA
18	1000189	1	NA	NA	1465	NA	NA	NA	NA	NA	NA	NA	NA	NA
19	1000190	1	NA	NA	1135	NA	NA	NA	NA	NA	NA	NA	NA	NA
20	1000205	1	NA	NA	1473	NA	NA	NA	NA	NA	NA	NA	NA	NA

Showing 1 to 22 of 502.599 entries, 178 total columns

A Raw UKB Subtable: Example

We extract from a ***.r.tab** file, columns (fields) have headers **f.datafieldnum.visit.arrayindx** i.e. 'f.20002.1.3'

[illegible]

A Raw UKB Subtable: Example

002.2.25	f.20002.2.26	f.20002.2.27	f.20002.2.28	f.20009.0.0	f.20009.0.1	f.20009.0.2	f.20009.0.3	f.20009.0.4	f.20009.0.5	f.20009.0.6	f.20009.0.7	f.20009.0.8
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	55 45630	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	23 00380	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	62 50000	62 500000	57 4247	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	66 30120	61 323500	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	57 90390	24 500000	66 9034	55 50000	NA	NA	NA	NA	NA
NA	NA	NA	NA	50 50000	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	38 48120	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	45 50000	53 500000	51 6724	54 67320	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	10 50000	28 695900	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	37 50000	37 500000	37 5000	37 50000	66 5000	40 5000	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	-1 00000	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	37 02460	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	61 07160	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	34 23480	NA	NA	NA	NA	NA	NA	NA	NA

Showing 1 to 23 of 502,599 entries. 178 total columns

Converted Table for conditions Asthma, Hay-Fever, Eczema including Age of Onset:

ID	Asthma_v0	Asthma_v1	Asthma_v2	Hayf_Rhin_v0	Hayf_Rhin_v1	Hayf_Rhin_v2	Eczema_v0	Eczema_v1	Eczema_v2	age_Asthma_v0	age_Asthma_v1	age_Asthma_v2
1	1000016	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
2	1000021	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
3	1000033	0	0	0	1	0	0	0	0	NA	NA	NA
4	1000049	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
5	1000057	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
6	1000065	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
7	1000070	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
8	1000082	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
9	1000098	0	NA	NA	1	NA	NA	0	NA	NA	NA	NA
10	1000103	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
11	1000111	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
12	1000126	1	NA	NA	0	NA	NA	0	NA	NA	53.5000	NA
13	1000134	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
14	1000142	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
15	1000155	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
16	1000167	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
17	1000178	1	NA	NA	0	NA	NA	0	NA	NA	-1.0000	NA
18	1000189	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
19	1000190	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
20	1000205	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA

Showing 1 to 22 of 502,599 entries, 19 total columns

Converted Table for conditions Asthma, Hay-Fever, Eczema including Age of Onset:

r1	Eczema_v2	age_Asthma_v0	age_Asthma_v1	age_Asthma_v2	age_Hayf_Rhin_v0	age_Hayf_Rhin_v1	age_Hayf_Rhin_v2	age_Eczema_v0	age_Eczema_v1	age_Eczema_v2
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0	NA	NA	NA	NA	44.5785	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	24.50000	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	53.5000	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	-1.0000	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Showing 1 to 22 of 502,599 entries, 19 total columns

UK Biobank Assessment Center → Touchscreen → Pain

biobank^{uk}

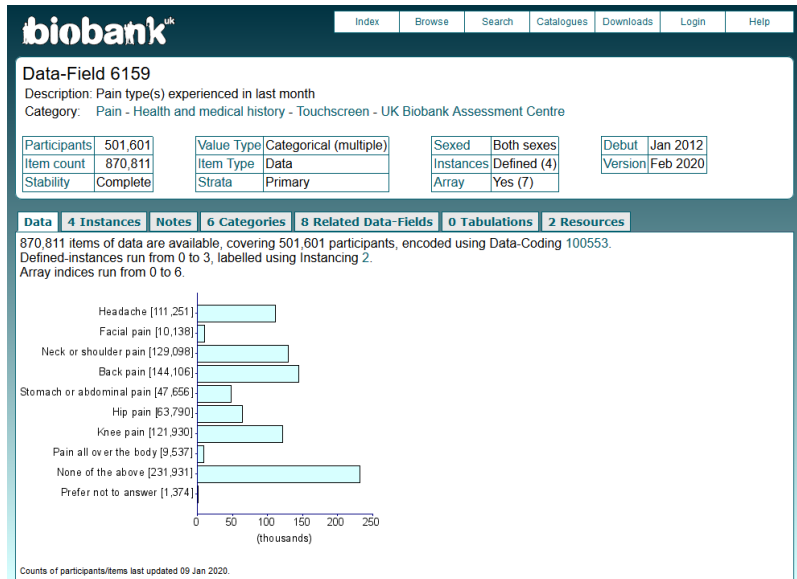
[Index](#) | [Browse](#) | [Search](#) | [Catalogues](#) | [Downloads](#) | [Login](#) | [Help](#)

Browse by Primary Category of Origin

Category	Items	
Population characteristics	35	Top Level
UK Biobank Assessment Centre	0	Level 1
Recruitment	20	
Touchscreen	0	Level 2
Sociodemographics	29	
Lifestyle and environment	155	Level 3
Early life factors	8	
Family history	20	
Psychosocial factors	48	
Health and medical history	0	
Eyesight	24	
Mouth	2	
General health	4	
Breathing	2	
Claudication and peripheral artery disease	10	
Pain	9	
Chest pain	4	
Cancer screening	4	
Operations	2	
Medical conditions	21	
Medication	11	
Hearing	9	
Sex-specific factors	34	
Verbal interview	33	
Physical measures	413	

Data-Field 6159: Data

Check: Data, Related Data-Fields, instancing 2, Data-Coding 100553, Notes



Data-Field 6159: Related Fields

biobank^{uk}

[Index](#) [Browse](#) [Search](#) [Catalogues](#) [Downloads](#) [Login](#) [Help](#)

Data-Field 6159

Description: Pain type(s) experienced in last month

Category: [Pain - Health and medical history](#) - [Touchscreen](#) - [UK Biobank Assessment Centre](#)

Participants	501,601	Value Type	Categorical (multiple)	Sexed	Both sexes	Debut	Jan 2012
Item count	870,811	Item Type	Data	Instances	Defined (4)	Version	Feb 2020
Stability	Complete	Strata	Primary	Array	Yes (7)		

[Data](#) [4 Instances](#) [Notes](#) [6 Categories](#) [8 Related Data-Fields](#) [0 Tabulations](#) [2 Resources](#)

Field ID	Description	Relationship
3571	Back pain for 3+ months	Field 3571 was collected from participants who indicated that in the last month they experienced back pain that interfered with their usual activities, as defined by their answers to Current Field
4067	Facial pains for 3+ months	Field 4067 was collected from participants who indicated that in the last month they experienced facial pain that interfered with their usual activities, as defined by their answers to Current Field
2956	General pain for 3+ months	Field 2956 was collected from participants who indicated that in the last month they experienced pain all over the body that interfered with their usual activities, as defined by their answers to Current Field
3799	Headaches for 3+ months	Field 3799 was collected from participants who indicated that in the last month they experienced headache that interfered with their usual activities, as defined by their answers to Current Field
3414	Hip pain for 3+ months	Field 3414 was collected from participants who indicated that in the last month they experienced hip pain that interfered with their usual activities, as defined by their answers to Current Field
3773	Knee pain for 3+ months	Field 3773 was collected from participants who indicated that in the last month they experienced knee pain that interfered with their usual activities, as defined by their answers to Current Field
3404	Neck/shoulder pain for 3+ months	Field 3404 was collected from participants who indicated that in the last month they experienced neck or shoulder pain that interfered with their usual activities, as defined by their answers to Current Field
3741	Stomach/abdominal pain for 3+ months	Field 3741 was collected from participants who indicated that in the last month they experienced stomach or abdominal pain that interfered with their usual activities, as defined by their answers to ~F61

Improving the health of future generations

Data-Field 6159: Data-Coding

biobank^{uk}

[Index](#)[Browse](#)[Search](#)[Catalogues](#)[Downloads](#)[Login](#)[Help](#)

Data-Coding 100553

Name: g22.i.answer.a.SY5.ans

Description: g22.i.answer.a.SY5.ans

This is a flat (unstructured) list which uses integers to represent categories or special values.

Coding can be downloaded here as a tab-separated file. [Download](#)

10 Categories

1 Data-Field

Coding	Meaning
1	Headache
2	Facial pain
3	Neck or shoulder pain
4	Back pain
5	Stomach or abdominal pain
6	Hip pain
7	Knee pain
8	Pain all over the body
-7	None of the above
-3	Prefer not to answer

Improving the health of future generations

A Raw UKB Subtable 6159 and Related Fields

missing ×															t6159 ×	Table_demogr_data.R ×		Set_of_functions.R ×		Table_condition_age_grp.R ×		frq_pain_vs_asthma.R ×		pain ×		conditions ×		t_cond_agegrp_bin ×		frq_pain_vs_cond ×		t20002 ×			
Filter																																			
field	f.6159.0.0	f.6159.0.1	f.6159.0.2	f.6159.0.3	f.6159.0.4	f.6159.0.5	f.6159.0.6	f.6159.1.0	f.6159.1.1	f.6159.1.2	f.6159.1.3	f.6159.1.4	f.6159.1.5	f.6159.1.6																					
1 1000016	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
2 1000021	1	3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
3 1000033	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
4 1000049	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
5 1000057	1	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
6 1000065	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
7 1000070	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
8 1000082	1	4	5	6	7	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
9 1000098	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
10 1000103	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
11 1000111	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
12 1000126	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
13 1000134	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
14 1000142	3	4	6	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
15 1000155	6	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
16 1000167	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
17 1000178	1	4	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
18 1000189	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
19 1000190	5	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					
20 1000205	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA																					

Showing 1 to 22 of 502,599 entries, 46 total columns

A Raw UKB Subtable 6159 and Related Fields

f.3571.1.0	f.3571.2.0	f.3741.0.0	f.3741.1.0	f.3741.2.0	f.3773.0.0	f.3773.1.0	f.3773.2.0	f.3799.0.0	f.3799.1.0	f.3799.2.0	f.4067.0.0	f.4067.1.0	f.4067.2.0
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA
NA	0	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	1	NA	NA	0	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	1	NA	NA	1	NA	NA	1	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	0	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA
NA	1	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	1	NA	NA	NA	NA	NA	NA	NA	NA

Showing 1 to 22 of 502,599 entries, 46 total columns

Converted Table for Pain and Duration:

ID	Prf_no_Anshr_v0	Non_Abve_v0	All_over_v0	Neck_Shldr_pn_v0	Hip_pn_v0	Back_pn_v0	Stom_Abdomn_pn_v0	Knee_pn_v0	Headch_v0	Face_pn_v0	Prf_no_Anshr_v1
1	1000016	0	1	0	0	0	0	0	0	0	NA
2	1000021	0	0	1	0	0	0	0	1	0	NA
3	1000033	0	1	0	0	0	0	0	0	0	0
4	1000049	0	1	0	0	0	0	0	0	0	NA
5	1000057	0	0	0	0	0	0	1	1	0	NA
6	1000065	0	1	0	0	0	0	0	0	0	NA
7	1000070	0	1	0	0	0	0	0	0	0	NA
8	1000082	0	0	0	1	1	1	1	1	0	NA
9	1000098	0	0	0	0	0	0	1	0	0	NA
10	1000103	0	1	0	0	0	0	0	0	0	NA
11	1000111	0	0	0	0	0	0	0	1	0	NA
12	1000126	0	1	0	0	0	0	0	0	0	NA
13	1000134	0	0	0	0	1	0	0	0	0	NA
14	1000142	0	0	1	1	1	0	1	0	0	NA
15	1000155	0	0	0	1	0	0	1	0	0	NA
16	1000167	0	1	0	0	0	0	0	0	0	NA
17	1000178	0	0	0	1	1	0	0	1	0	NA
18	1000189	0	0	0	0	0	0	1	0	0	NA
19	1000190	0	0	0	0	0	1	1	0	0	NA
20	1000205	0	0	0	0	0	0	1	0	0	NA

Showing 1 to 22 of 502.599 entries. 31 total columns

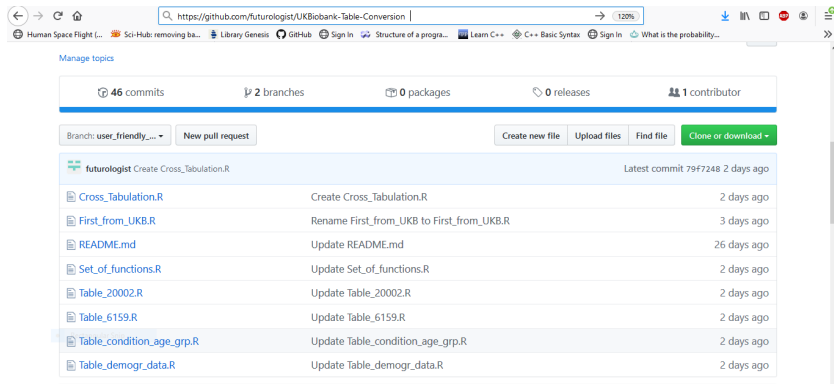
Converted Table for Pain and Duration:

Face_pn_v2	All_over_3+m_v0	Neck_Shldr_pn_3+m_v0	Hip_pn_3+m_v0	Back_pn_3+m_v0	Stom_Abdomn_pn_3+m_v0	Knee_pn_3+m_v0	Headch_3+m_v0	Face_pn_3+m_v0	All_over_3
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	0	NA	NA	NA	NA	1	NA	NA
0	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	1	0	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	1	1	1	1	1	NA	NA
NA	NA	NA	NA	NA	NA	1	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	0	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	1	NA	NA	NA	NA	NA
NA	NA	1	1	1	NA	1	NA	NA	NA
NA	NA	NA	1	NA	NA	1	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	0	0	NA	NA	0	NA	NA
NA	NA	NA	NA	NA	NA	1	NA	NA	NA
NA	NA	NA	NA	NA	1	1	NA	NA	NA
NA	NA	NA	NA	NA	NA	1	NA	NA	NA

Showing 1 to 22 of 502.599 entries. 55 total columns

R Code for Table Conversion:

Important: The code extracts from ***.r.tab** files, columns (fields) have headers **f.datafieldnum.visit.arrayindx** i.e. **f.6159.1.3**



The screenshot shows a GitHub repository page for 'futurolgist/UKBiobank-Table-Conversion'. The repository has 46 commits, 2 branches, 0 packages, 0 releases, and 1 contributor. The current branch is 'user_friendly_...'. The repository contains several R files, with the latest commit being 'Create Cross_Tabulation.R' 2 days ago. The files listed are:

File	Commit Message	Time
Cross_Tabulation.R	Create Cross_Tabulation.R	2 days ago
First_from_UKB.R	Rename First_from_UKB to First_from_UKB.R	3 days ago
README.md	Update README.md	26 days ago
Set_of_functions.R	Update Set_of_functions.R	2 days ago
Table_20002.R	Update Table_20002.R	2 days ago
Table_6159.R	Update Table_6159.R	2 days ago
Table_condition_age_grp.R	Update Table_condition_age_grp.R	2 days ago
Table_demogr_data.R	Update Table_demogr_data.R	2 days ago

Converts some subtables of the original raw UKB data into analysis friendly data tables

Instruction:

Step 1: Use script **First_from_UKB.R** to extract a raw subtable from the big raw UKB data file

The next scripts import and use the file of R functions:

Set_of_functions.R

Step 2: After the raw subtable is being extracted

2.1. Field 20002 and related: Use **Table_20002.R**

2.2. Field 6159 and related: Use **Table_6159.R**

2.3. Fields of demographic data (e.g. Age, Gender, Principal Components, etc): Use **Table_demogr_data.R**

Step 3: Further conversions:

3.1. A medical condition from field 20002 together with categorization for Age of Onset: Use **Table_condition_age_grp.R**

3.2. Cross-tabulation of two tables: Use **Cross_Tabulation.R**

First_from_UKB.R

```
#!/usr/bin/env Rscript

library(dplyr)
library(data.table)

extract_UKB_subtable <- function(data_base, list_of_codes){
  headers <- names(data_base)
  pos <- c(1)
  for(code in list_of_codes){
    pos <- c(pos, grep(code, headers))
  }
  subt <- data_base[, ..pos]
  subt <- subt[order(subt[, 'f.eid']), ]
  subt
}

# For a set of fields that need to be extracted from the central UKB data table
# list a vector of strings, each string in the format "f.[Field_number].", for each field.
# Skip the first column, id column "f.eid", it is automatically included

##### YOUR INPUT GOES HERE: #####

list_of_codes <- c("f.6159.", "f.2956.", "f.3404.", "f.3414.", "f.3571.", "f.3741.", "f.3773.", "f.3799.", "f.4067.")
#list_of_codes <- c("f.135.", "f.20002.", "f.20009.")
#list_of_codes <- c("f.137.", "f.20003.")

filepath_in <- "/mnt/nfs/backup/data/uk_biobank/ukb22741.r.tab" #string
filepath_out <- "/home/ndimit2/Asthma_and_Pain/output_data_ph1/ukb6159.txt"
#####

data_base <- fread(filepath_in)
subtable <- extract_UKB_subtable(data_base, list_of_codes)
write.table(subtable, filepath_out, append = FALSE, sep = "\t", quote = FALSE, col.names=TRUE, row.names=FALSE)

#To run this from the UNIX command line, type and execute: chmod +x table_extraction.r
# ./table_extraction.r world
```

The Result is a Raw UKB Subtable:

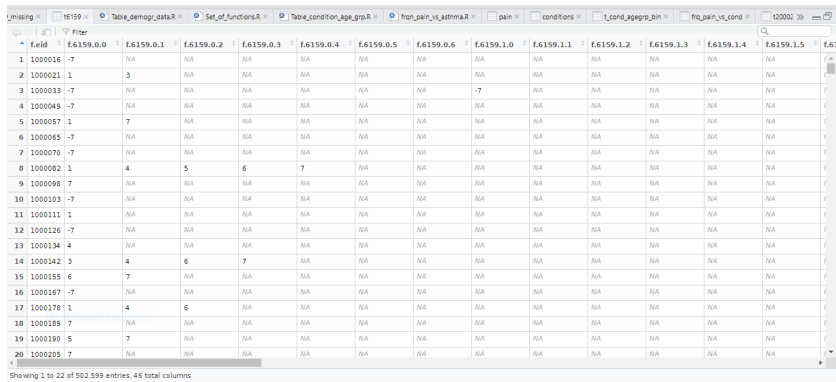
Here: Related Data-Fields 135 (number of self-reported conditions)
20002 (self-reported conditions) and 20009 (age of onset)

	f.field	f.135.0.0	f.135.1.0	f.135.2.0	f.20002.0.0	f.20002.0.1	f.20002.0.2	f.20002.0.3	f.20002.0.4	f.20002.0.5	f.20002.0.6	f.20002.0.7	f.20002.0.8	f.20002.0.9
1	1000016	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
2	1000021	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
3	1000033	1	5	0	1537	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
4	1000049	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
5	1000057	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
6	1000065	1	N/A	N/A	1462	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
7	1000070	3	N/A	N/A	1065	1473	1465	N/A	N/A	N/A	N/A	N/A	N/A	N/A
8	1000082	2	N/A	N/A	1278	1309	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
9	1000098	4	N/A	N/A	1465	1387	1529	1666	N/A	N/A	N/A	N/A	N/A	N/A
10	1000103	1	N/A	N/A	1075	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
11	1000111	1	N/A	N/A	1426	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
12	1000126	4	N/A	N/A	1065	1111	1138	1286	N/A	N/A	N/A	N/A	N/A	N/A
13	1000134	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
14	1000142	2	N/A	N/A	1412	1311	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
15	1000155	6	N/A	N/A	1075	1074	1065	1473	1465	1078	N/A	N/A	N/A	N/A
16	1000167	0	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
17	1000178	1	N/A	N/A	1111	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
18	1000189	1	N/A	N/A	1465	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
19	1000190	1	N/A	N/A	1135	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
20	1000205	1	N/A	N/A	1473	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Showing 1 to 22 of 502.599 entries. 178 total columns

The Result is a Raw UKB Subtable:

Here: Data-Field 6159 (Pain) and related fields of duration of pain



The screenshot shows a data table with 22 columns and 20 rows. The columns are labeled as follows: field, f.6159.0.0, f.6159.0.1, f.6159.0.2, f.6159.0.3, f.6159.0.4, f.6159.0.5, f.6159.0.6, f.6159.1.0, f.6159.1.1, f.6159.1.2, f.6159.1.3, f.6159.1.4, f.6159.1.5, and f.6159.1.6. The rows contain numerical data, with some cells containing 'NA' or specific values like -7, 1, 3, 4, 5, 6, 7. The table is displayed in a software interface with a menu bar at the top and a status bar at the bottom.

	field	f.6159.0.0	f.6159.0.1	f.6159.0.2	f.6159.0.3	f.6159.0.4	f.6159.0.5	f.6159.0.6	f.6159.1.0	f.6159.1.1	f.6159.1.2	f.6159.1.3	f.6159.1.4	f.6159.1.5	f.6159.1.6
1	1000016	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2	1000021	1	3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	1000033	-7	NA	NA	NA	NA	NA	NA	-7	NA	NA	NA	NA	NA	NA
4	1000049	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
5	1000057	1	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6	1000065	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
7	1000070	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8	1000082	1	4	5	6	7	NA	NA	NA	NA	NA	NA	NA	NA	NA
9	1000098	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	1000103	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	1000111	1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
12	1000126	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
13	1000134	4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
14	1000142	3	4	6	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
15	1000155	6	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
16	1000167	-7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
17	1000178	1	4	6	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
18	1000189	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
19	1000190	5	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
20	1000205	7	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Showing 1 to 22 of 502,599 entries, 46 total columns

Table_20002.R

```
source("C:\\MY_FOLDERS\\Asthma_and_Pain\\R_code\\Table_creation\\Set_of_functions.R")

##### INPUT FOR CONDITIONS #####
filepath20002_in <- "C:\\MY_FOLDERS\\Asthma_and_Pain\\output_data_ph1\\ukb20002.txt"

list_of_conditions <- c(1111, 1387, 1452)

list_of_labels <- c('Asthma', 'Hayf_Rhin', 'Eczema')

filepath20002_out <- "C:\\MY_FOLDERS\\Asthma_and_Pain\\output_data_ph2\\asth_rhin_ecz.txt"
#####

##### NOT TO BE CHANGED #####
n_visits <- 3
l_array <- 29 # there are 29 fields for conditions per visit
start_pos <- 5 # the columns with codes for conditions start with column 5 and end with columns 5+3*29-1 = 91
#####

##### EXECUTE #####
t20002 <- fread(filepath20002_in) #502599

t20002_bin <- build_cond_and_age_diag_table(t20002,
                                             list_of_conditions,
                                             list_of_labels,
                                             n_visits,
                                             l_array,
                                             start_pos)

names(t20002_bin)[1] <- 'ID'

t20002_bin_reord <- group_by_visit(t20002_bin, list_of_conditions, n_visits)
t20002_cond_age <- group_cond_age(t20002_bin, list_of_conditions, n_visits)

write.table(t20002_bin,|
            filepath20002_out,
```

Converted Table for conditions Asthma, Hay-Fever, Eczema including Age of Onset:

ID	Asthma_v0	Asthma_v1	Asthma_v2	Hayf_Rhin_v0	Hayf_Rhin_v1	Hayf_Rhin_v2	Eczema_v0	Eczema_v1	Eczema_v2	age_Asthma_v0	age_Asthma_v1	age_Asthma_v2
1	1000016	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
2	1000021	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
3	1000033	0	0	0	1	0	0	0	0	NA	NA	NA
4	1000049	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
5	1000057	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
6	1000065	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
7	1000070	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
8	1000082	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
9	1000098	0	NA	NA	1	NA	NA	0	NA	NA	NA	NA
10	1000103	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
11	1000111	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
12	1000126	1	NA	NA	0	NA	NA	0	NA	NA	53.5000	NA
13	1000134	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
14	1000142	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
15	1000155	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
16	1000167	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
17	1000178	1	NA	NA	0	NA	NA	0	NA	NA	-1.0000	NA
18	1000189	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
19	1000190	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA
20	1000205	0	NA	NA	0	NA	NA	0	NA	NA	NA	NA

Showing 1 to 22 of 502,599 entries, 19 total columns

Converted Table for conditions Asthma, Hay-Fever, Eczema including Age of Onset:

r1	Eczema_v2	age_Asthma_v0	age_Asthma_v1	age_Asthma_v2	age_Hayf_Rhin_v0	age_Hayf_Rhin_v1	age_Hayf_Rhin_v2	age_Eczema_v0	age_Eczema_v1	age_Eczema_v2
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	0	NA	NA	NA	NA	44.5785	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	24.50000	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	53.5000	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	-1.0000	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Showing 1 to 22 of 502,599 entries, 19 total columns

Table_6159.R

```
source("C:\\MY_FOLDERS\\Asthma_and_Pain\\R_code\\Table_creation\\Set_of_functions.R")
##### INPUTS FOR PAIN SITES #####
filepath6159_in <- "C:\\MY_FOLDERS\\Asthma_and_Pain\\Output_data_ph1\\ukb6159.txt"

# you can change the lables that follow, but preservet their order:
list_of_labels <- c('Prf_no_Anscr', 'Non_Abve', 'All_over',
                   'Neck_Shldr_pn', 'Hip_pn', 'Back_pn', 'Stom_Abdmn_pn', 'Knee_pn', 'Headch', 'Face_pn')
#####

##### INPUTS FOR PAIN DURATION #####
# you can change the lables that follow, but preservet their order:
labels <- c('All_over_3+m', 'Neck_shldr_pn_3+m',
            'Hip_pn_3+m', 'Back_pn_3+m', 'Stom_Abdmn_pn_3+m',
            'Knee_pn_3+m', 'Headch_3+m', 'Face_pn_3+m')

filepath6159_out <- "C:\\MY_FOLDERS\\Asthma_and_Pain\\Output_data_ph2\\UKB_Pain_Duration.txt"
#####

##### NOT TO BE CHANGED #####
n_visits <- 3 # The number of visits can change to 4 when the fourth visit becomes available and is incuded in the UKB table
list_of_sites <- c(-3, -7, 8, 3, 6, 4, 5, 7, 1, 2) # Do not change.
l_array <- 7 # there are 7 fields for pain siates per visit
start_pos <- 2 # the columns with codes for pain sites start with column 2 and end with column 2+3*7-1 = 22

#The order of the conditions and the lables is based on the order of the fields from the file First_from_UKB.R
#They were:]
#list_of_codes <- c("f.6159.", "f.2956.", "f.3404.", "f.3414.", "f.3571.", "f.3741.", "f.3773.", "f.3799.", "f.4067.")
#####

##### NOT TO BE CHANGED #####
#fields <- c("f.3799.", "f.4067.", "f.3404.", "f.3571.", "f.3741.", "f.3414.", "f.3773.", "f.2956.")
fields <- c("f.2956.", "f.3404.", "f.3414.", "f.3571.", "f.3741.", "f.3773.", "f.3799.", "f.4067.")
arrays_length <- c(1, 1, 1, 1, 1, 1, 1, 1)
instances <- 3 * arrays_length
#####
```

Converted Table for Pain and Duration:

ID	Prf_no_Anshr_v0	Non_Abve_v0	All_over_v0	Neck_Shldr_pn_v0	Hip_pn_v0	Back_pn_v0	Stom_Abdomn_pn_v0	Knee_pn_v0	Headch_v0	Face_pn_v0	Prf_no_Anshr_v1
1	1000016	0	1	0	0	0	0	0	0	0	NA
2	1000021	0	0	1	0	0	0	0	1	0	NA
3	1000033	0	1	0	0	0	0	0	0	0	0
4	1000049	0	1	0	0	0	0	0	0	0	NA
5	1000057	0	0	0	0	0	0	1	1	0	NA
6	1000065	0	1	0	0	0	0	0	0	0	NA
7	1000070	0	1	0	0	0	0	0	0	0	NA
8	1000082	0	0	0	1	1	1	1	1	0	NA
9	1000098	0	0	0	0	0	0	1	0	0	NA
10	1000103	0	1	0	0	0	0	0	0	0	NA
11	1000111	0	0	0	0	0	0	0	1	0	NA
12	1000126	0	1	0	0	0	0	0	0	0	NA
13	1000134	0	0	0	0	1	0	0	0	0	NA
14	1000142	0	0	1	1	1	0	1	0	0	NA
15	1000155	0	0	0	1	0	0	1	0	0	NA
16	1000167	0	1	0	0	0	0	0	0	0	NA
17	1000178	0	0	0	1	1	0	0	1	0	NA
18	1000189	0	0	0	0	0	0	1	0	0	NA
19	1000190	0	0	0	0	0	1	1	0	0	NA
20	1000205	0	0	0	0	0	0	1	0	0	NA

Showing 1 to 22 of 502,599 entries. 31 total columns

Converted Table for Pain and Duration:

Face_pn_v2	All_over_3+m_v0	Neck_Shldr_pn_3+m_v0	Hip_pn_3+m_v0	Back_pn_3+m_v0	Stom_Abdomn_pn_3+m_v0	Knee_pn_3+m_v0	Headch_3+m_v0	Face_pn_3+m_v0	All_over_3
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	0	NA	NA	NA	NA	1	NA	NA
0	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	1	0	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	1	1	1	1	1	NA	NA
NA	NA	NA	NA	NA	NA	1	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	0	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	1	NA	NA	NA	NA	NA
NA	NA	1	1	1	NA	1	NA	NA	NA
NA	NA	NA	1	NA	NA	1	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	0	0	NA	NA	0	NA	NA
NA	NA	NA	NA	NA	NA	1	NA	NA	NA
NA	NA	NA	NA	NA	1	1	NA	NA	NA
NA	NA	NA	NA	NA	NA	1	NA	NA	NA

Showing 1 to 22 of 502.599 entries. 55 total columns

Table_demogr_data.R

```
source("C:\\MY_FOLDERS\\Asthma_and_Pain\\R_code\\Table_creation\\Set_of_functions.R")

#If you would like to extract additional data, such as demographic and principle components from UKB, use this script
#The tables are extracted "as they are" columns are selected as desired
#When selecting a subset of the fields, make sure the order of the fields is in ascending order
#(as it is in the raw UKB table) and the array lengths, the instances and the labels correspond to the respective fields
# also make sure, that the fields you are choosing are present in the raw UKB data table at pathfile_in

##### INPUTS FOR ADDITIONAL DATA, E.G. DEMOGRAPHIC INFO #####
pathfile_in <- "C:\\MY_FOLDERS\\Asthma_and_Pain\\output_data_ph1\\ukb_demogr_geno_info.txt"

fields <- c('f.31.', 'f.21000.', 'f.21003.', 'f.22001.', 'f.22006.', 'f.22009.', 'f.22010.', 'f.22018.')
arrays_length <- c(1, 1, 1, 1, 1, 40, 1, 1)
instances <- c(1, 3, 3, 1, 1, 1, 1, 1)
labels <- c("Sex", "Ethnic_backgr", "Age_at_visit", "Genetic_sex",
            "Gen_ethnic_grp", "PC", "Geno_analys_exclns",
            "Relat_exclns")

#fields <- c(21000, 22006, 22009)
#arrays_length <- c(1, 1, 40)
#instances <- c(3, 1, 1)
#labels <- c("Ethnic_backgr", "Gen_ethnic_grp", "PC")

filepath_out <- "C:\\MY_FOLDERS\\Asthma_and_Pain\\output_data_ph2\\asth_rhin_ecz_demogr.txt"
#####

##### EXECUTE #####
t_demogr_geno_info <- fread(pathfile_in) #502599
t_demogr_geno_info <- extract_subtable(t_demogr_geno_info, fields)
names(t_demogr_geno_info) <- relabel(t_demogr_geno_info, fields, array_length, instances, labels)
names(t_demogr_geno_info)[1] <- 'ID'

asth_rhin_ecz_demogr <- left_join(t20002_bin, t_demogr_geno_info, by = "ID")

write.table(asth_rhin_ecz_demogr, filepath_out,
            append = FALSE, sep = "\\t", quote = FALSE, col.names=TRUE, row.names=FALSE)
```

Demographic Table:

ID	Sex	Ethnic_backgr_v0	Ethnic_backgr_v1	Ethnic_backgr_v2	Genetic_sex	Gen_ethnic_grp	PC_1	PC_2	PC_3	PC_4	PC_5	PC_6	PC_7
1	1000016	1	1001	N/A	N/A	1	-14.971300	6.569230	-1.3365300	0.92670900	-5.1387900	0.71295100	1.00131
2	1000021	1	1003	N/A	N/A	1	-14.386900	4.848020	-1.4494900	0.16317500	-12.7189000	4.05945000	-6.6349
3	1000033	1	4002	6	4002	1	N/A	404.239000	78.346700	-7.3963800	5.06768000	1.3545800	1.26281000
4	1000049	0	1001	N/A	N/A	0	-14.186200	5.049800	1.1151700	1.44953000	3.7704100	-1.02480000	-1.1049
5	1000057	1	1001	N/A	N/A	1	-11.732400	2.347860	-2.1769300	-0.77984700	-1.5601300	0.30691700	0.79560
6	1000065	1	1001	N/A	N/A	1	-11.669800	3.402110	0.1570150	-0.68362400	-9.2771500	-0.54041600	1.30480
7	1000070	1	1001	N/A	N/A	1	-13.510500	6.019790	-2.2976800	3.54059000	0.3880490	-0.07259120	-1.7648
8	1000082	0	1001	N/A	N/A	0	-12.255400	3.367310	-2.0331000	-3.20088000	-7.8891200	-1.47467000	0.65190
9	1000098	0	1001	N/A	N/A	0	-12.380700	6.652540	-0.3084850	3.61791000	0.3283630	2.03677000	0.86090
10	1000103	1	1001	N/A	N/A	1	-13.652900	4.977300	-2.0118900	6.04367000	10.5052000	-0.21671700	-1.0284
11	1000111	1	1001	N/A	N/A	1	-11.124800	2.267640	0.3657850	0.48853700	-1.5978400	1.34270000	1.75740
12	1000126	1	1001	N/A	N/A	1	-13.801400	1.992730	-0.3710420	5.02652000	5.0188100	-3.45319000	-0.8898
13	1000134	1	1001	N/A	N/A	1	-12.784500	2.322710	-1.4030400	3.80499000	16.1347000	-1.82734000	1.18040
14	1000142	0	1001	N/A	N/A	0	-12.883700	5.586510	-0.1269750	-2.31585000	-5.9124100	-0.50693400	-1.5391
15	1000155	1	1001	N/A	N/A	1	-9.397750	1.280260	-2.3414600	-1.88726000	1.7767700	-0.88299100	3.77990
16	1000167	1	1001	N/A	N/A	1	-12.984000	2.344330	-3.4956900	2.87668000	-6.1619000	-0.83405200	0.29260
17	1000178	1	1001	N/A	N/A	0	-12.691700	3.537740	-5.2217800	3.09525000	2.5255200	-1.69996000	0.50970
18	1000189	1	1001	N/A	N/A	1	-9.596010	1.870230	-3.2655500	0.33014600	-4.2363800	-1.20747000	0.31200
19	1000190	0	1001	N/A	N/A	0	-15.294400	3.606250	-3.3196700	7.13003000	15.6512000	-0.13457200	2.68750
20	1000205	1	1001	N/A	N/A	1	21.153400	-11.274600	28.8372000	-81.36310000	15.1426000	-5.69029000	-25.020

Showing 1 to 22 of 502,599 entries. 47 total columns

Table_condition_age_grp.R

```
source("C:\\MY_FOLDERS\\Asthma_and_Pain\\R_code\\Table_creation\\Set_of_functions.R")

# THIS SCRIPT USES THE RESULTS FROM THE SCRIPT 'Table_20002.R', SO YOU MAY HAVE TO RUN 'Table_20002.R' FIRST.
# FOR THIS SCRIPT TO WORK PROPERLY, YOU HAVE TO MAKE SURE THE HEALTH CONDITION LABEL (E.G. "Asthma")
# PROVIDED HERE MATCHES THE CORRESPONDING LABEL IN THE
# TABLE(S) EXTRACTED WITH THE SCRIPT 'Table_20002.R'.

##### NOT TO BE CHANGED #####
n_visits <- 3 # The number of visits you go up to 4 when updates of the database is available
l_array <- 29 # there are 29 fields for conditions per visit
start_pos <- 5 # the columns with codes for conditions start with column 5 and end with columns 5+3*29-1 = 91
#####

##### INPUTS FOR CONDITION SELECTION WITH AGE GROUPS #####
# Choose one of the tables, generated by the script 'Table_20002.R' :

t_cond <- t2000p2_bin

# Refer to the headers of either table t20002_bin, t20002_bin_reord or t20002_cond_age,
# or alternatively, the script 'Table_20002.R' and the choose the label from the provided there
condition <- "Asthma"

age_groups <- list(c(0,18), c(18,40), c(40, 120)) # Provide your own age group intervals
age_group_lbls <- c('Age_6', 'Age_4', 'Age_5') # Provide your own labels for the age groups above
filepath_out <- "C:\\MY_FOLDERS\\Asthma_and_Pain\\Output_data_ph2\\asthma_age_groups.txt"

##### EXECUTE #####
t_cond_agegrp_bin <- build_cond_agegrp_bin(t_cond, condition, age_groups, age_group_lbls, n_visits, keep_NA = TRUE)

write.table(t_cond_agegrp_bin,
            filepath_out,
            append = FALSE, sep = "\\t", quote = FALSE, col.names=TRUE, row.names=FALSE)
```

Condition with Age of Onset Groups:

ID	Asthma_v0	Asthma_Age_6_v0	Asthma_Age_4_v0	Asthma_Age_5_v0	Asthma_v1	Asthma_Age_6_v1	Asthma_Age_4_v1	Asthma_Age_5_v1	Asthma_v2	Asthma_Age_6_v2
1 1000016	0	0	0	0	NA	NA	NA	NA	NA	NA
2 1000021	0	0	0	0	NA	NA	NA	NA	NA	NA
3 1000033	0	0	0	0	0	0	0	0	0	0
4 1000049	0	0	0	0	NA	NA	NA	NA	NA	NA
5 1000057	0	0	0	0	NA	NA	NA	NA	NA	NA
6 1000065	0	0	0	0	NA	NA	NA	NA	NA	NA
7 1000070	0	0	0	0	NA	NA	NA	NA	NA	NA
8 1000082	0	0	0	0	NA	NA	NA	NA	NA	NA
9 1000098	0	0	0	0	NA	NA	NA	NA	NA	NA
10 1000103	0	0	0	0	NA	NA	NA	NA	NA	NA
11 1000111	0	0	0	0	NA	NA	NA	NA	NA	NA
12 1000126	1	NA	NA	1	NA	NA	NA	NA	NA	NA
13 1000134	0	0	0	0	NA	NA	NA	NA	NA	NA
14 1000142	0	0	0	0	NA	NA	NA	NA	NA	NA
15 1000155	0	0	0	0	NA	NA	NA	NA	NA	NA
16 1000167	0	0	0	0	NA	NA	NA	NA	NA	NA
17 1000178	1	NA	NA	NA	NA	NA	NA	NA	NA	NA
18 1000189	0	0	0	0	NA	NA	NA	NA	NA	NA
19 1000190	0	0	0	0	NA	NA	NA	NA	NA	NA
20 1000205	0	0	0	0	NA	NA	NA	NA	NA	NA

Showing 1 to 22 of 502,559 entries, 13 total columns

Cross_Tabulation.R

```
library(dplyr)
library(data.table)

# THIS SCRIPT USES THE REUSLTS FROM THE SCRIPTS 'Table_condition_age_grp.R' and 'Table_6159.R',
# OR ALTERNATIVELY, FROM YOUR OWN SCRIPTS
# SO YOU MAY HAVE TO RUN THESE TWO SCRIPTS FIRST.

create_frequency_table <- function(pain, conditions){
  id <- intersect(select(pain, 'ID'), select(conditions, 'ID'))
  pain_loc <- left_join(id, pain, by='ID')
  conditions_loc <- left_join(id, conditions, by='ID')
  pain_loc <- pain_loc[, (2:length(pain[1,]))]
  conditions_loc <- conditions_loc[, (2:length(conditions_loc[1,]))]
  pain_loc[is.na(pain_loc)] <- 0
  conditions_loc[is.na( conditions_loc)] <- 0

  D <- data.matrix(pain_loc)
  P <- data.matrix(conditions_loc)
  R <- t(D) %>% P
  R <- as.data.table(R, keep.rownames = TRUE)
  return(R)
}

#####
##### PROVIDE OUPUT LOCATION #####
filepath_out <- "C:\\MY_FOLDERS\\Asthma_and_Pain\\output_data_ph2\\cross_table_pain_vs_cond.txt"
#####

conditions <- t_cond_agegrp_bin
pain <- t_pain

cross_pain_vs_cond <- create_frequency_table(pain, conditions)
```


Cross Table Pain Sites vs a Medical Condition:

rn	Missing_20002_v0	Controls_v0	Asthma_v0	Asthma_Age_6_v0	Asthma_Age_4_v0	Asthma_Age_5_v0	Missing_20002_v1	Controls_v1	Asthma_v1	Asthma_Age_6_v1
1	Missing_6159_v0	391	1634	166	25	42	71	2169	20	0
2	Non_Abve_v0	143	178258	18735	6163	4748	6459	188231	8040	865
3	All_over_v0	21	7208	1616	283	446	746	8620	187	38
4	Neck_Shldr_pn_v0	131	100353	16535	4178	4460	6533	112687	3721	611
5	Hip_pn_v0	57	47748	8498	1835	2102	3799	54279	1723	301
6	Back_pn_v0	143	111915	18052	4564	4834	7157	125371	4108	631
7	Stom_Abdom_pn_v0	57	36881	7057	1772	2162	2538	42410	1308	277
8	Knee_pn_v0	109	93061	15025	3690	3826	6218	104191	3444	560
9	Headch_v0	125	87955	14908	3927	4691	5012	99157	3267	564
10	Face_pn_v0	15	7807	1521	331	452	592	8977	310	56
11	Missing_6159_v1	856	429103	56685	16009	15191	21025	482262	3650	732
12	Non_Abve_v1	1	8921	965	315	247	332	2	8878	1007
13	All_over_v1	0	184	34	6	7	20	0	181	37
14	Neck_Shldr_pn_v1	2	3229	504	124	147	196	1	3192	542
15	Hip_pn_v1	2	1815	294	72	82	120	0	1800	311
16	Back_pn_v1	3	3571	543	147	154	205	0	3544	573
17	Stom_Abdom_pn_v1	0	1006	177	39	62	66	0	997	186
18	Knee_pn_v1	2	3431	513	131	141	205	0	3423	523
19	Headch_v1	1	2196	350	81	128	111	0	2174	373
20	Face_pn_v1	0	214	43	9	11	17	0	211	46

Showing 1 to 22 of 30 entries. 19 total columns