



OXFORD



**POPULATION
HEALTH**

Introduction to the UK Biobank

Xiaonan Liu, Jennifer Collister, Lei Clifton
Nuffield Department of Population Health
Oxford University

Part I: Introduction to UKB

Part II: Data Downloads

What's in UKB

- The go-to website is the [“Browse”](#) website.

Browse by Primary Category of Origin

Category	Items	Top Level
+ Population characteristics	35	Top Level
+ Assessment centre	3939	Level 1
+ Biological samples	978	Level 2
+ Genomics	271	Level 2
+ Online follow-up	1107	Level 2
+ Additional exposures	366	Level 3
+ Health-related outcomes	2646	Level 3
		Level 4

“Items” shows the number of data fields within each category.

Browse by Primary Category of Origin

Category	Items	Top Level
- Population characteristics	0	Top Level
+ Baseline characteristics	31	Level 1
Ongoing characteristics	4	
- Assessment centre	0	Level 2
+ Recruitment	13	Level 2
+ Touchscreen	396	Level 3
+ Cognitive function	121	Level 3
+ Verbal interview	37	Level 3
+ Physical measures	264	Level 4
+ Eye measures	333	Level 4
+ Imaging	2691	
+ Biological sampling	10	
+ Procedural metrics	74	

You can view the data available at different levels.

Common data fields

- Participant ID (FID eid – short for “encoded ID”)
- Sex ([FID 31](#))
- Date of birth

We use these variables all the time!

(approximated by 15th of [FID 52](#) “Month of birth” and [FID 34](#) “Year of birth”)

- Date of recruitment (i.e. baseline) ([FID 53](#))
- Self-reported ethnicity ([FID 21000](#))
- Recruitment centres ([FID 54](#))

There’s also genetic ethnicity
([FID 22006](#))

Data fields – things to be aware of!

- Instances and Arrays
- Fields that aren't available at baseline
- Fields that aren't available for the whole population
- Derived fields
- Special/Missing data codes

These will be covered in more detail
in the following slides

Instances and Arrays

- Instances: How many visits participants have this measurement performed?

4 visits in UKB,
coded from 0-3

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+)

- Arrays: How many multiple measurements at one visit per person? (e.g. Blood pressure, etc)

Different data fields might have different arrays.

UKB Reference: Section 5 of "[Intro to UKB Showcase](#)".

Instances and Arrays

Data-Field 6154

Description: Medication for pain relief, constipation, heartburn

Category: Assessment centre ▶ Touchscreen ▶ Health and medical history ▶ Medication

Participants	498,700
Item count	684,466
Stability	Complete

Value Type	Categorical (multiple)
Item Type	Data
Strata	Primary

Sexed	Both sexes
Instances	Defined (4)
Array	Yes (6)

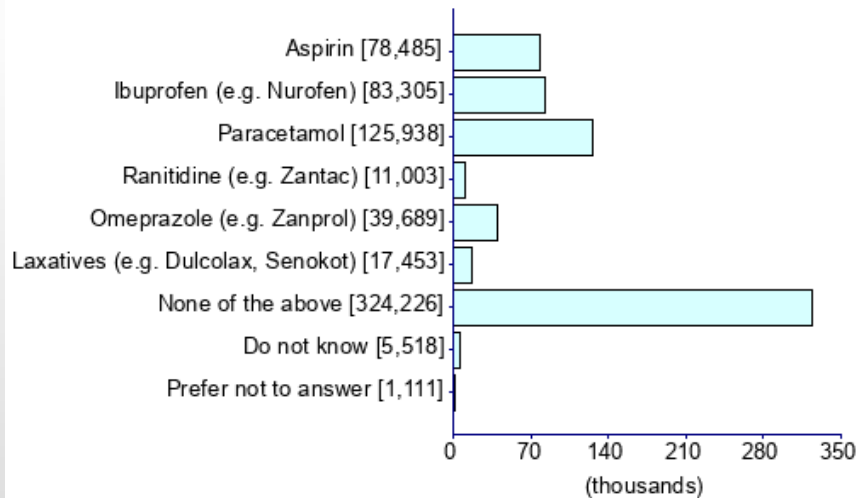
Debut	Jan 2012
Version	Nov 2022
Cost Tier	d1 o1 s1

Data 4 Instances Notes 6 Categories 2 Related Data-Fields 2 Resources

684,466 items of data are available, covering 498,700 participants, encoded using Data-Coding 100628.

Defined-instances run from 0 to 3, labelled using Instancing 2.

Array indices run from 0 to 5.



Dummy data, showing format of column name: FieldID-Instance.Array

ID	6154-0.0	6154-0.1	...	6154-0.5
1	1	2	NA	NA
2	3	NA	NA	NA
3

Coding

Meaning

1	Aspirin
2	Ibuprofen (e.g. Nurofen)
3	Paracetamol
4	Ranitidine (e.g. Zantac)
5	Omeprazole (e.g. Zanprol)
6	Laxatives (e.g. Dulcolax, Senokot)
-7	None of the above
-1	Do not know
-3	Prefer not to answer

Good practice: check field description on Showcase

Fields unavailable at baseline

- [Imaging, Category 100003](#)

Imaging	0
Abdominal MRI	4
Kidney MRI	4
Liver MRI	6
Pancreas MRI	4
Abdominal composition	28
Abdominal organ composition	15

- [Online follow-up, Category 100089](#)

Online follow-up	0
Cognitive function online	56
Diet by 24-hour recall	473
Digestive health	54
Experience of pain	129
Food (and other) preferences	153
Mental health	142
Work environment	100

Data-Field 20243

Description: Kidney Imaging - T1 ShMOLLI - DICOM

Category: Assessment centre ▶ Imaging ▶ Abdc

Participants	6,559	Value Type	Text
Item count	6,559	Item Type	Bulk
Stability	Accruing	Strata	Primary

Data	2 Instances	Notes	4 Categories	0 I
------	-------------	-------	--------------	-----

Instance 2 : Imaging visit (2014+)

4,529 participants, 4,529 items

Data not meaningful for statistical summary

Instance 3 : First repeat imaging visit (2019+)

2,030 participants, 2,030 items

Data not meaningful for statistical summary

Check "Instances" tab of data fields to check whether data are available at baseline.

Fields that aren't available for the whole population

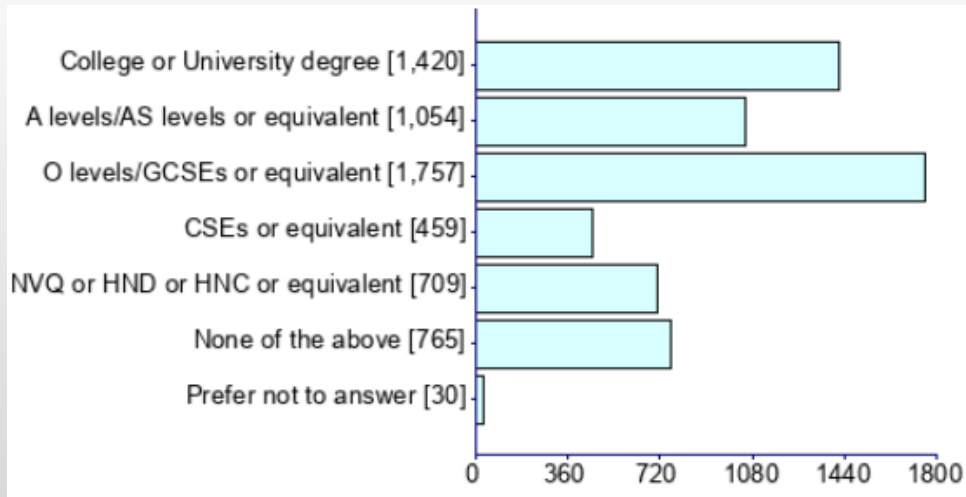
- Introduced partway through recruitment (e.g. [FID 4559](#))

Data	4 Instances	Notes	5 Categories	0 Related Data-Fields	2 Resources
------	-------------	-------	--------------	-----------------------	-------------

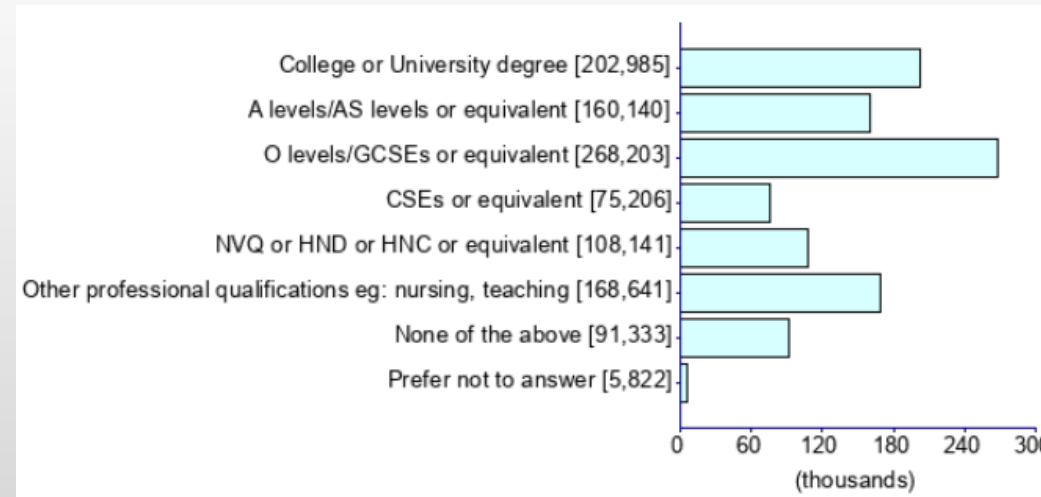
ACE touchscreen question "In general how satisfied are you with your FAMILY RELATIONSHIPS?"

Question was introduced part way through fieldwork in April 2009.

- Pilot study (asked/measured slightly different) (e.g. Education [FID 10722](#))



Pilot field: FID 10722



Non-pilot field: FID 6138

Our approach: merge if possible.

Fields that aren't available for the whole population

- Asked to subgroups only (e.g. Alcohol intake)

[FID 1558](#): Alc intake freq

Daily or almost daily [117,011]
Three or four times a week [139,750]
Once or twice a week [152,531]
One to three times a month [66,039]
Special occasions only [67,238]
Never [46,660]
Prefer not to answer [629]

[FID 4407](#): Average monthly red wine intake

Data 4 Instances **Notes** Categories 1 Related Data-Fields 2 Resources

ACE touchscreen question "In an average MONTH, how many glasses of RED wine would you drink? (There are six glasses in an average bottle)"

The following checks were performed:

- If answer < 0 then rejected
- If answer > 250 then rejected
- If answer > 10 then participant asked to confirm

If the participant activated the Help button they were shown the message:

Please include sparkling red wine here.

Field 4407 was collected from participants who indicated they drink alcohol on special occasions or one to three times a month, as defined by their answers to Field 1558

- Different by gender
 - E.g. [ID 6177](#) and [ID 6153](#) both ask about medications, but is split by gender, so women can be additionally asked about hormonal contraceptives and HRT. We merge them into one variable.

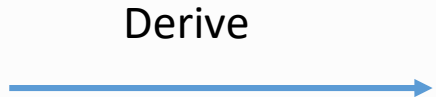
Derived fields

- Some fields are derived by UKB

Our suggestion: explicitly state that these variables were derived (i.e. not directly measured or asked)

Data-Field 1239
Description: Current tobacco smoking

Data-Field 1249
Description: Past tobacco smoking



Data-Field 20116
Description: Smoking status
Category: Assessment centre ▶ Touchscreen ▶ Lifestyle and

Participants	501,516
Item count	587,921
Stability	Complete

Value Type	Categorical (single)
Item Type	Data
Strata	Derived

Data-Field 26217
Description: Enhanced PRS for body mass index (BMI)
Category: Genomics ▶ Polygenic Risk Scores ▶ Enhanced PRS

Participants	104,600
Item count	104,600
Stability	Complete

Value Type	Continuous, relative risk
Item Type	Data
Strata	Derived

Data-Field 21001
Description: Body mass index (BMI)
Category: Assessment centre ▶ Physical measures ▶ Anthro

Participants	499,405
Item count	584,330
Stability	Complete

Value Type	Continuous, Kg/m2
Item Type	Data
Strata	Derived

Special/Missing codes

- Special/Missing codes (especially for continuous variable)

Data-Field 2704

Description: Years since last cervical smear test

Category: [Assessment centre](#) ▶ [Touchscreen](#) ▶ [Sex-specific factors](#) ▶ [Female-specific factors](#)

Participants	265,887	Value Type	Integer, years	Sexed	Females only	Debut	Jan 2012
Item count	309,230	Item Type	Data	Instances	Defined (4)	Version	Nov 2022
Stability	Complete	Strata	Primary	Array	No	Cost Tier	d1 o1 s1

Data	4 Instances	Notes	5 Categories	1 Related Data-Fields	2 Resources
------	-------------	-------	--------------	-----------------------	-------------

ACE touchscreen question "How many years ago was your last cervical smear test?"

The following checks were performed:

- If answer < 0 then rejected
- If answer > Participants age - 10 years then rejected
- If answer > 15 then participant asked to confirm

If the participant activated the Help button they were shown the message:

If you are unsure, please provide an estimate or select Do not know.

Field 2704 was collected from women who indicated that they had had a cervical smear test, as defined by their answers to Field 2694

Coding 100569 defines 3 special values:

- -10 represents "Less than a year ago"
- -1 represents "Do not know"
- -3 represents "Prefer not to answer"

Our approach

- For characteristics table, we show the missing categories (e.g. income).
- For analysis, we bundle them as NA.

Summary on data fields

Data-Field 6153
 Description: Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones
 Category: Assessment centre ▶ Touchscreen ▶ Health and medical history ▶ Medication

Participants	271,319	Value Type	Categorical (multiple)	Sexed	Females only	Debut	Jan 2012
Item count	347,298	Item Type	Data	Instances	Defined (4)	Version	Nov 2022
Stability	Complete	Strata	Primary	Array	Yes (4)	Cost Tier	d1 o1 s1

Things to be aware of	Suggested To-dos
Instances and Arrays	Check the variable description banner.
Fields that aren't available at baseline	Check the "Instances" tab
Fields that aren't available for the whole population	Check the "Notes" tab
Derived fields	Check the "Notes" tab and be clear of writing in manuscript.
Special/Missing codes	Check the "Notes" tab.

- Always check UKB Showcase for your variables.
- Every scenario can be project specific, so you need your own tailored solution.

Useful resources

- How to find data fields
 - Do a search: <https://biobank.ndph.ox.ac.uk/showcase/search.cgi>
 - Search UKB related papers on your project
- Withdrawal list
 - UKB updates researchers via email
- The [Schema](#) for the Data Showcase contains lots of useful info mappings for the data
 - Download schema → Individual csv file for coding
 - e.g. [ID 20002](#) “Non cancer illness code” uses the hierarchical [Data Coding 6](#) to encode the various medical conditions participants could self-report in the verbal interview

Part I: Introduction to UKB

Part II: Data Downloads

Basket naming conventions

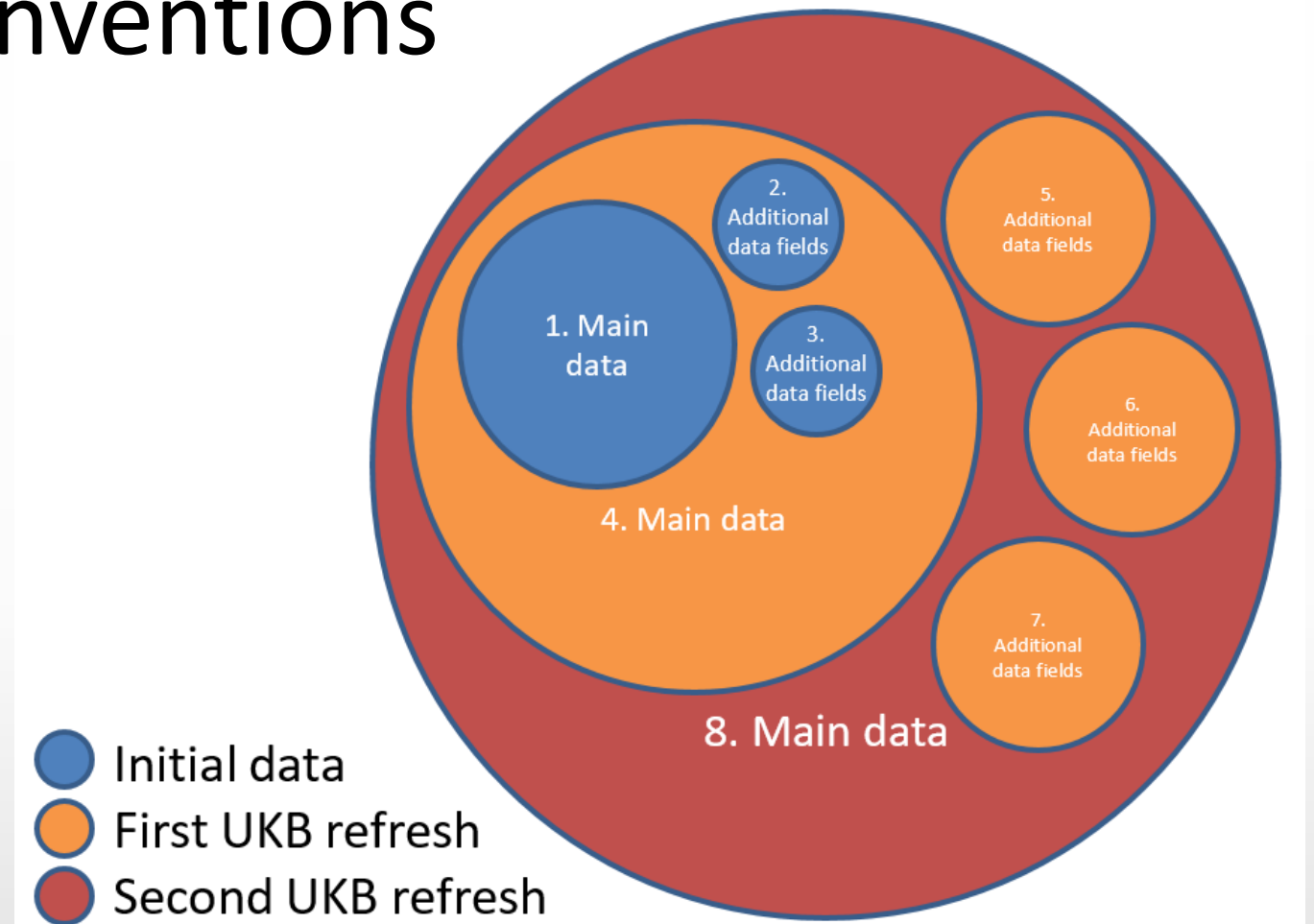
- **Keep a download log.**

- **Principles:**

- Having all our data in one basket makes it easier to see what variables we have access to

BUT

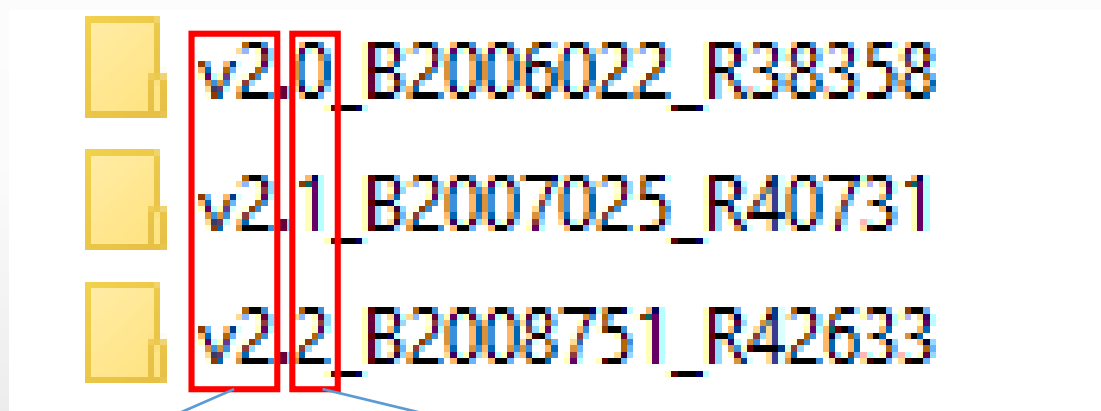
- When we realise we want a couple of extra variables, we don't want to have to download ALL the data again.



UKB “[Managing Baskets](#)” guide has instructions on creating and merging UKB data baskets.

Our download naming conventions

- Our naming convention is
v<version number>_B<Basket ID>_R<Run ID>
- The Basket and Run IDs for the data download will be in the email from UKB



MAJOR version when
UKB provides
refreshed data

MINOR version when we make
a new small basket containing
a few additional fields

Censoring dates

- Censoring dates: [Data providers and dates of data availability](#)

Hospital Admissions (Inpatients)	Data Provider	International Classification of Diseases (ICD)		Classification of Interventions and Procedures (OPCS)		Period of data currently available	Censoring date
		ICD9	ICD10	OPCS3	OPCS4		
Hospital Episode Statistics for England (HES)	NHS Digital		1997 onwards		1997 onwards	1997 onwards, with critical care data from 2011	30 September 2021 *
Scottish Morbidity Record (SMR)	Information and Statistics Division (ISD), Scotland	1981 - 1996	1996 onwards	1977 - 1988	1989 onwards	1981 onwards	31 July 2021 **
Patient Episode Database for Wales (PEDW)	Secure Anonymised Information Linkage (SAIL), Wales		1999 onwards		1999 onwards	1998 onwards	28 February 2018***

Be aware: the censoring dates on Showcase get updated at each UKB data refresh!

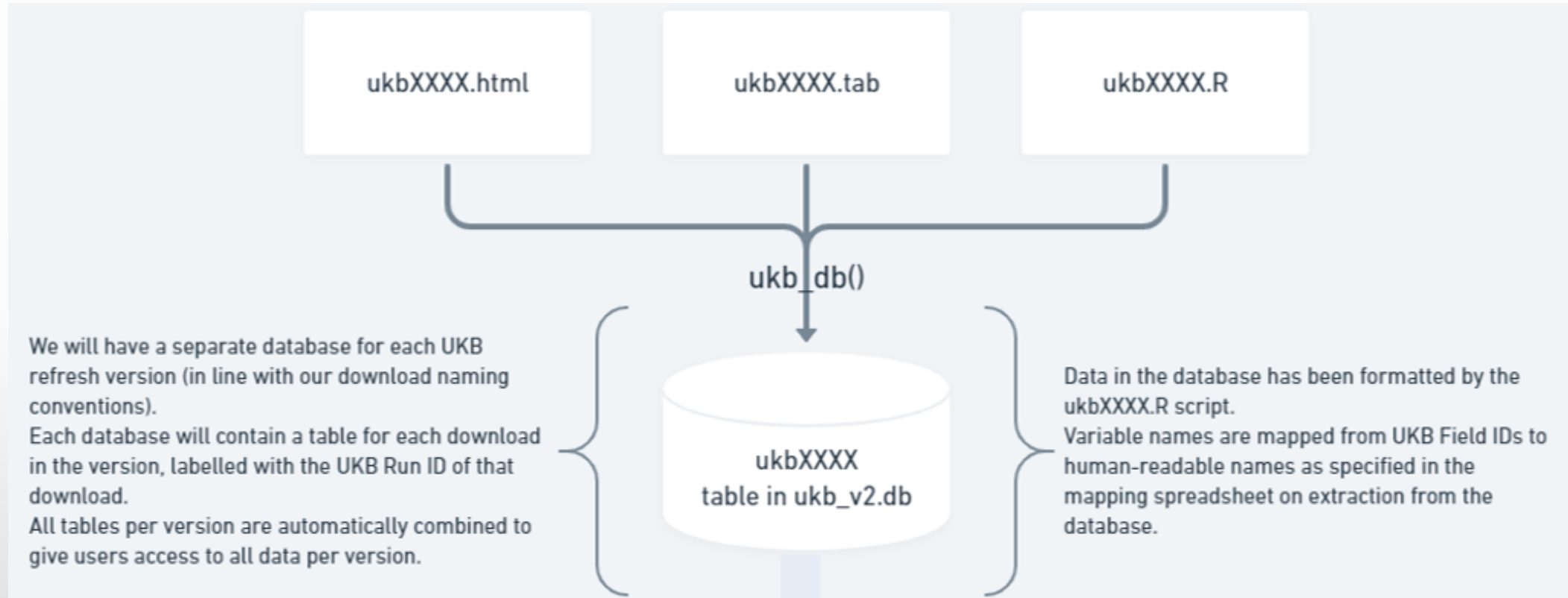
```
censoring.yml - Notepad
File Edit Format View Help
# Provide date in format %Y-%m-%d
England: "2020-09-30"
Scotland: "2020-08-31"
Wales: "2018-02-28"
```

Our approach: record the censoring date in a static file at each download of our data.

Database



- Our solution to memory issue: use [duckdb](#) database



- How to generate the database: Jennifer's code [here](#) with accompanying documentation [here](#) (described in following slides)

Database approach

- Split the huge .tab file into several smaller chunks, each containing only around 25k rows (participants)
- Read these chunks into memory one at a time, and write to a database
- This initial setup step only needs to be done once for a given data download
 - Using a high performance computing cluster (eg the BMRC) this should only take a couple of hours
- Data can then be extracted quickly from the database for analysis

Code for generating the database

- Download the UKB data and convert it to R format to get ukbXXX.html, ukbXXX.tab, ukbXXX.R (see Section 2.3 in [Data Access Guide](#)).

More information in “[Setting up the database](#)” guide.

- Split the .tab file into smaller chunks using `split -l 25000 -d --additional-suffix=.tab ukbXXX.tab ukbXXX_`

- Run [data toDB.R](#)

```
ukb_db(fileset = download_runID,
       path = file_path,
       dbname = db_name,
       chunks = chunks,
       mapping = mapping,
       stata = stata
)
```

Prefix of the .html, .tab, and .R file

Path to the folder containing the ukbXXXXX files

Name of the database to write to

Number of chunks to read

Renaming sheet

Set this to TRUE if you want to be able to extract data from the database using Stata

Accompanying documentation for the database

- Extract data from db in R: see [here](#).

```
DB_extract(  
  extract_cols,  
  db = config$data$database,  
  name_map = config$cleaning$renaming,  
  withdrawals = config$cleaning$withdrawals  
)
```

Specify the columns you wish to extract, file path of database, file path of renaming file, and file path of withdrawals list.

- Extract data from db in Python or Stata: see [here](#).
- Cautionary note: duckdb package is still under development, which means that unfortunately new versions are often not backwards compatible. This means a database written under one version of duckdb cannot be read by a later version.

Our solution: using "[renv](#)" for version control of packages

Renaming UKB variables

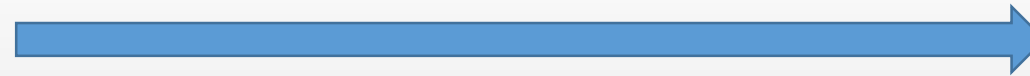
Field_ID	Field_Description	NewVarName
47	Hand grip strength (right)	HGS_R
48	Waist circumference	BSM_Waist
52	Month of birth	BaC_BirthMonth
53	Date of attending assessment centre	Rec_DateAssess
54	UK Biobank assessment centre	Rec_AssessCentre
120	Birth weight known	Vel_BirthWtKnown

Eg

BaC = Baseline Characteristics

Vel = Verbal Interview

ID	47-0.0
1	...
2	...



Use renaming.csv

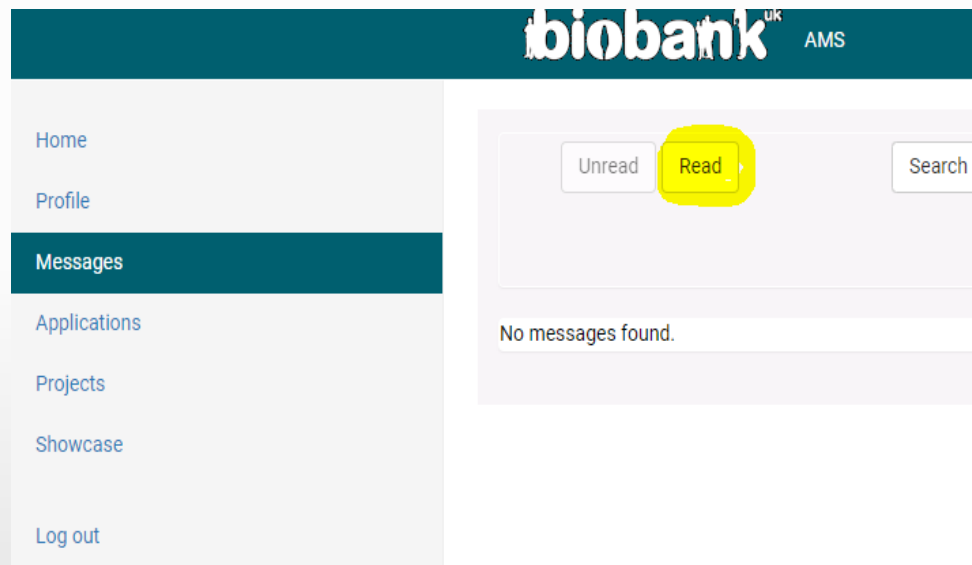
ID	HGS_R-0.0
1	...
2	...

UKB Raw data:
Format of column name is
FieldID-Instance.Array

UKB processed data:
Replace FieldID with
NewVarName

Messages from UKB

- Messages in Access Management System (AMS)



The message will be marked as “Read”, as soon as the 1st person has read it.

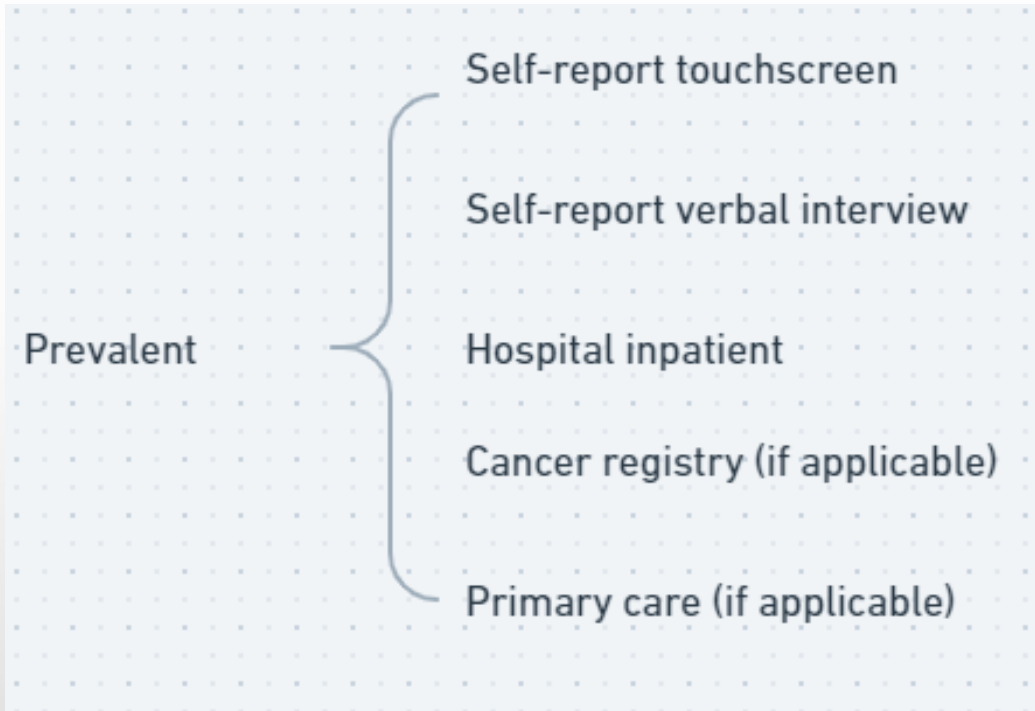
- **Suggestion:** the 1st person to read the message take a screenshot and send to others. This saves everyone’s time and avoids confusion!

Health outcome 1/4

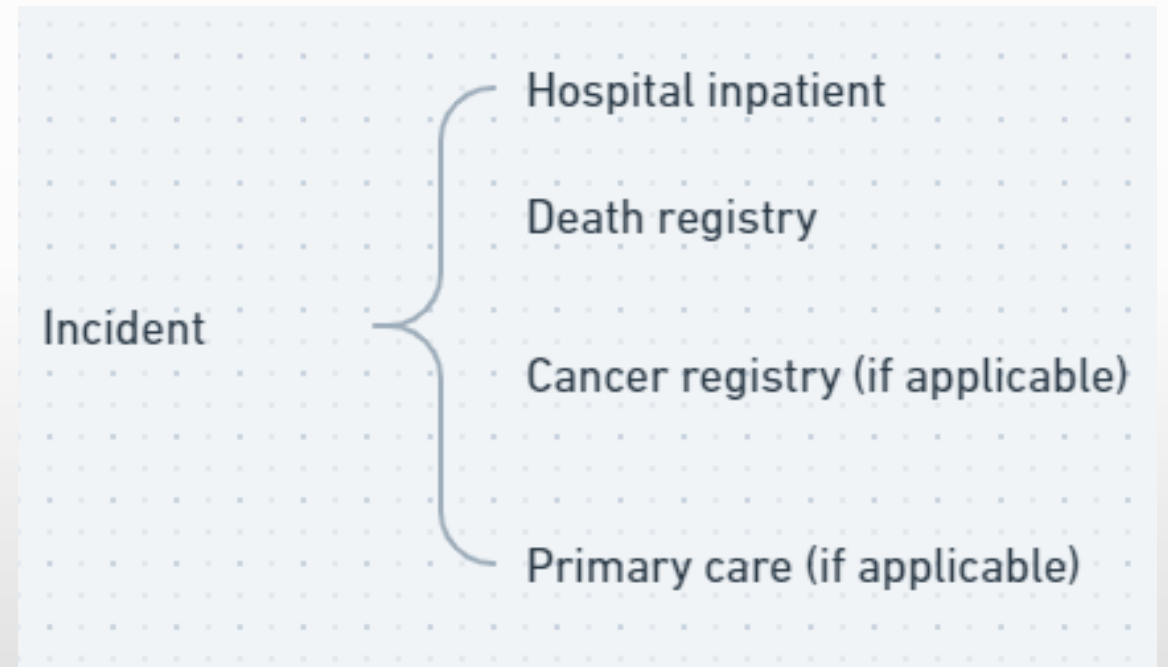
Resources	UKB Category or Field ID	Available platforms
Self-report touchscreen questionnaires	Medical conditions (Category 100044), Medication (Category 100045)	Showcase
Self-report verbal interview	Medical conditions (Category 100074), Medications (Category 100075), Operations (Category 100076)	Showcase
Hospital inpatient	ICD 10 (FID 41270, 41280), ICD 9 (FID 41271, 41281), OPCS 4 (FID 41272, 41282)	Showcase, Data portal
Death registry	ICD 10 (FID 40001, 40002, 40000)	Showcase, Data portal
Cancer registry	ICD 10 (FID 40006) ICD 9 (FID 40013) Date of cancer diagnosis (FID 40005)	Showcase
Primary care records	Record-level access (Category 3001)	Data portal

Health outcome 2/4

- Prevalent vs Incident



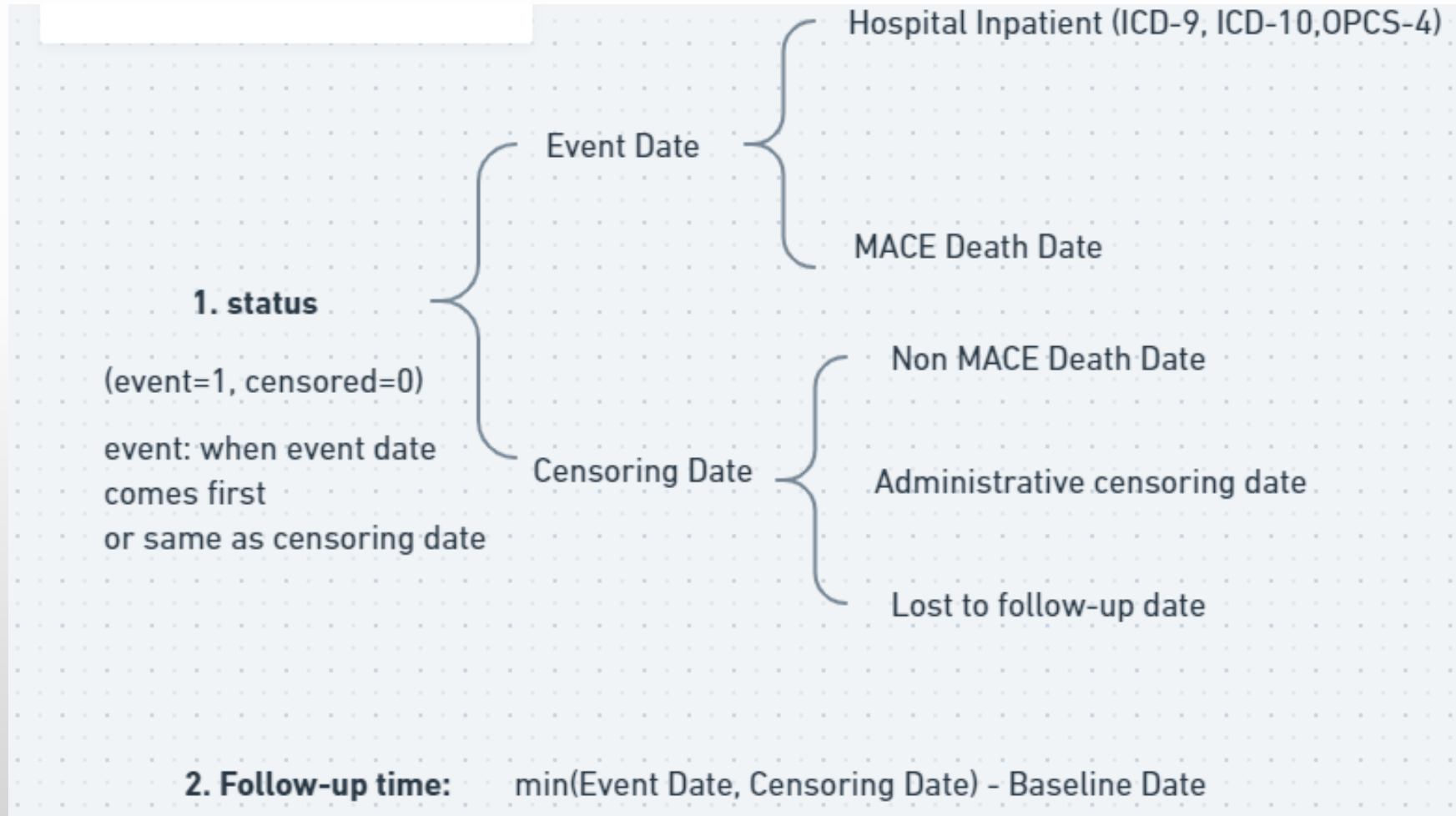
Prevalent cases: no need to incorporate death registry.



Incident cases: we usually don't incorporate self-report data

Health outcome 3/4

- How to identify incident cases using multiple resources: e.g. Major Adverse Cardiovascular Event (MACE) outcome

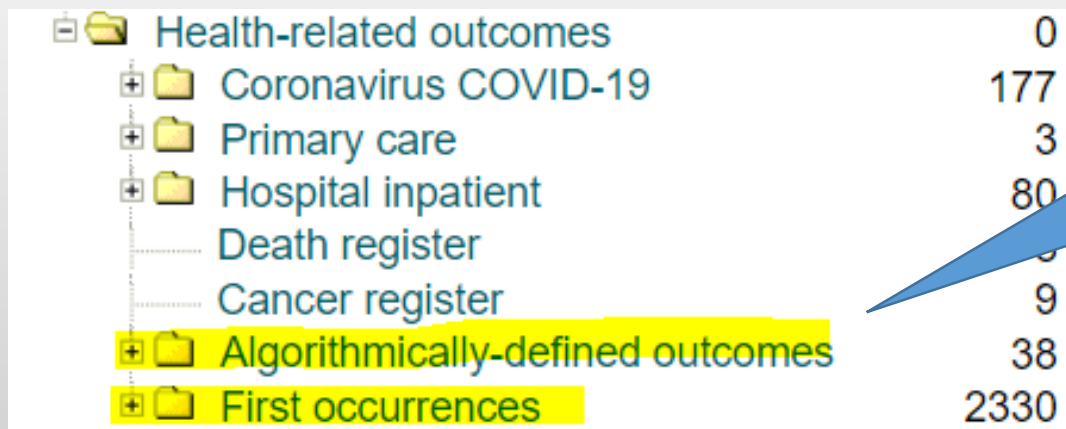


Health outcome 4/4

- Our recommendation: use record-level if possible, because it is updated more frequently than Showcase.

	Summary-level data	Record-level data
Data	Records the distinct diagnoses and procedures and dates these were first recorded	All instances of diagnoses and procedures
How to download	Download from Showcase (the basket system)	Download from Data Portal

- Other categories



Health-related outcomes	0
+ Coronavirus COVID-19	177
+ Primary care	3
+ Hospital inpatient	80
Death register	0
Cancer register	9
+ Algorithmically-defined outcomes	38
+ First occurrences	2330

We don't really use these data fields but the clinical codes could be useful.

UKB "[Health Outcomes Overview](#)"

Genetic data

- Genotyping data: [Category 263](#)
- Description of genetic data types: [Resource 531](#)
- Instructions for downloading genotype data: [Resource 668](#)
 - Use the shared institution copy if possible, as it saves resources.
 - We access the institutional copy on the BMRC cluster (you need to contact the BMRC team for access)
 - You still need to download the unique link files for your application (.fam, .sample)
- Exome/Whole genome: Only available via the Research Analysis Platform (RAP).

UKB Resources

- Main page of resources: [Index](#) -> [Essential Information](#) -> [Accessing UK Biobank Data](#) has links to
 - [Data Access Guide](#)
 - [RAP documentation](#)
 - [Summary of data available: Data providers and dates of data availability](#) is particularly useful, because it has updated censoring dates when new data is released.
 - [Data Dictionary of Showcase Fields \(csv\)](#)
- Other links
 - [Showcase guide](#)
 - [Managing Baskets](#) (under the [Resource Catalogue](#) as [Resource 747](#))
 - [Synthetic dataset](#) – UK Biobank has made available a fictional dataset of the same size and constitution to their real data, to allow large scale system testing.

This is not the renaming.csv