UK Biobank: from downloading data to extracting variables

A step-by-step guide

Lei Clifton

1. Introduction

1.1 Motivation

The UK Biobank (UKB) "showcase" website provides good instructions on handling its datasets, but it is not always obvious where to start, due to the scale and complexity of the resource. I often find myself forgetting the details of past data curation tasks, so have written this step-by-step guide as an aide memoire to help myself and others. It shows you the first step of handling UKB data, from downloading data to extracting variables.

Disclaimer: this guide is only intended as informal and unofficial document, based on our personal experience with UKB data. Our research group (Translational Epidemiology Unit, TEU) will not be responsible for any consequences arising from the use of this guide.

1.2 Resources

You will need to download the following resources from the UKB showcase website:

https://biobank.ctsu.ox.ac.uk/showcase/

- Go to: UKB Showcase website -> Downloads -> 6 file handlers (ukbmd5, ukbconv, ukbunpack, ubkfetch, ukblink, ukbgene) and 1 miscellaneous utility (encoding.ukb). Download these six "file handlers", which are individual tools for downloading UKB data..
- Go to: UKB Showcase website -> Essential information -> Accessing your data -> Accessing data guide. This will show a pdf file, "Accessing_UKB_data_v2.0.pdf", which contains instructions on how to use the six file handlers. Save this pdf for future use.
- Go to: UKB Showcase website -> Essential information -> Requesting data and using the UK Biobank showcase -> Data dictionary: all fields available in UK Biobank (spreadsheet format). This option will automatically save a "Data_Dictionary_Showcase.csv" file to your PC, which is useful for specifying which variables to extract in Step 4 (see below).

The following resource is also useful, although not directly used in this document:

- Go to: UKB Showcase website -> Essential information -> Requesting data and using the UK Biobank showcase -> User guide: getting started with Showcase. This opens a pdf file "ShowcasUserGuide.pdf", which tells you how to look for your variables - save it for future use.

1.3 Before you start

You will need to log into the UKB website and download your UKB dataset, following p.9-10 in "Accessing_UKB_data_v2.0.pdf". Steps 1-5 (next section) assume that you have downloaded your UKB dataset with a filename similar to "*ukb26505.enc*".

NB I have used the Windows operating system for all tasks in this document. The guide does not deal with bulk (eg. image or genetic) data. Another guide is planned on how to use the "ukbgene" utility, which should be run on Linux (not Windows).

2. A step-by-step guide: converting the downloaded data to SAS (R, or Stata) format

Steps 1-3 run the downloaded file handlers (ukbmd5 etc.) in a Windows (ie. "DOS") command window, following the instructions given on p.11-17 in "Accessing_UKB_data_v2.0.pdf".

Steps 4-5 are not in that pdf file.

I have found that the six downloaded file handlers function correctly when run from the J drive (which is the NDPH network drive for personal files), but not when run from the K drive (the NDPH network drive for work). This does not matter, because Step 5 will save the labelled dataset in your specified folder on the K drive.

2.1 Step 1: validating the download (ukbmd5)

Create a folder called "Data" on your J drive. Put your downloaded file handlers and "*ukb26505.enc*" dataset into this folder.

Opened a command window by typing "cmd" in the Windows "Start" menu (i.e., after pressing the "Windows" button). Type "cd J:\Data" at the command line.

Type "ukbmd5" as shown below, to run that file handler. The returned MD5 should be the same as the MD5 that is stated in an automated email from UKB to you. Your command line and its output should look like this:

J:\Data>ukbmd5 ukb26505.enc
UKBiobank ukb_md5_win (c) CTSU. Compiled Dec 9 2013 10:31:44.
Input file: "ukb26505.enc" opened
Size=2465512, MD5= (the unique numbers in the email from UKB to you)

This step is the equivalent of p.11 in "Accessing_UKB_data_v2.0.pdf".

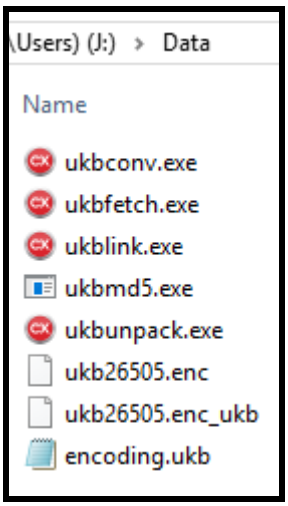
2.2 Step 2: decrypting the dataset (ukbunpack)

Run "ukbunpack", as shown below. Your command line should look like this:

J:\Data>ukbunpack ukb26505.enc keyvalue

The *keyvalue* you need to enter is a long string of letters and numbers with a filename like "k54321.key", given to you via email from UKB. Open this file in Notepad, copy the long string, and then right-click to paste it into the DOS window.

This step is the equivalent of p.12-13 in "Accessing_UKB_data_v2.0.pdf". It will create a new file named ""*ukb26505.enc_ukb*" in the same folder. Now your J:\Data folder should look like this:



2.3 Step 3: converting the dataset into SAS, R, or Stata format (ukbconv)

Run "ubkconv", as shown below. I usually run the following four commands; see p.13-17 of "Accessing_UKB_data_v2.0.pdf" for more options.

```
J:\Data>ukbconv ukb26505.enc_ukb docs
J:\Data>ukbconv ukb26505.enc_ukb sas
J:\Data>ukbconv ukb26505.enc_ukb r
J:\Data>ukbconv ukb26505.enc_ukb stata
```

If your dataset is particularly large, it will take time to run these commands! For example, my dataset has 470 distinct fields (3300 variables), and it takes "ukbconv" 9 hours (on a standard NDPH desktop in 2020) to convert it into the SAS format of size 9.7 GB.

2.4 Step 4: (optional) extracting specified fields

As noted above, steps 4 and 5 are not in "Accessing_UKB_data_v2.0.pdf".

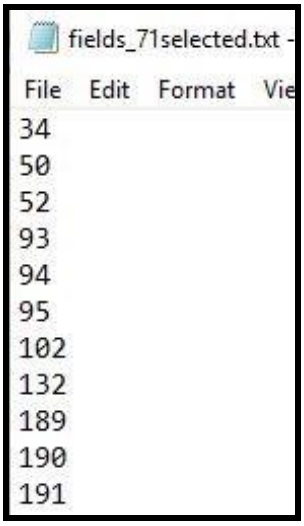
Your converted dataset produced in step 3 above can be large (e.g., more than 10 GB). You might want to extract some fields of particular interest, and save the extracted variables in a smaller dataset. You can do this using the "vlookup" (vertical look up) function in Excel, as described below.

The "Data_Dictionary_Showcase.csv" downloaded from the UKB Showcase website is useful for you to specify which fields you need. Column B ("FieldID") contains numbers "3, 4, 5, …" that the "ukbconv" command needs, while Column C ("Field") contains text labels that describe the fields.

The "vlookup" function in Excel allows you to specify your selected fields in Column C (a textual description of that field). The vlookup function will then extract the corresponding "FieldID" in Column B for you. This website has good instructions on how to use "vlookup":

https://support.office.com/en-us/article/vlookup-function-0bbc8083-26fe-4963-8ab8-93a18ad188a1

I have named the resulting txt file "*fields_71selected.txt*". This example file contains 71 rows indicating my chosen fields, with the first few lines looking like this (each of which are the FieldIDs from Column B):



I then run the following command to convert my 71 selected fields into SAS, R, and Stata format.

```
ukbconv ukb26505.enc_ukb sas -ifields_71selected.txt
ukbconv ukb26505.enc_ukb r  -ifields_71selected.txt
ukbconv ukb26505.enc_ukb stata  -ifields_71selected.txt
```

Note that there is no space between "*-i*" and "*fields_71selected.txt*" in the command option above.

2.5 Step 5: labelling variables (in SAS)

It is quicker to test your code on a small dataset instead of the complete (very large) dataset. Here I will use the dataset containing my selected 71 fields (254 variables) produced in Step 4 to illustrate how to label the variables in SAS.

The "ukbconv" command line with "sas" option in step 4 above will produce the following three files for me:

- ukb26505.sd2: SAS dataset containing my selected 71 fields, not ready for using yet.
- ukb26505.sas: I will slightly modify this SAS code, and save it as "ukb26505_LC.sas".
- ukb26505.log: a small log file.

The "ukb26505_LC.sas" will add value labels (ie. textual descriptions) to the variables in ukb26505.sd2, and save the labelled variables in the ".sas7bdat" format, such that are ready to use in SAS. My first modification of the code provided by UKB is at the beginning of the file where the unlabelled raw dataset "`raw_zbnzxi`" is read, as shown below:

```
 5  *LC: I renamed ukb26505.sas to ukb_71fields.sas.
 6    I aslo renamed dataset ukb26505.sd2 to ukb_71fields.sd2. ;
 7  dm "log; clear; "; /* Clear log window. Open a log file */
 8  filename fzbnzxi 'J:\Data\ukb_71fields.sd2';  /* my new file name */
 9
10  *LC: I specified my libname, so that the data are saved in my savedata library,
11      instead of SAS default work library;
12  libname savedata "K:\TEU\CancerPRS\1stData_12Feb2019\SelectedConvSAS_71fields";
13
14 data savedata.raw_zbnzxi; /*LC: I specfied my library savedata here*/
15    infile fzbnzxi RECFM=V LRECL=1960;
16    input n_eid 8. n_34_0_0 5. n_50_0_0 13. n_50_1_0 13. n_50_2_0 13. n_52_0_0 3.
```

The code as provided by UKB then uses "proc format" to define value labels (miaa, mpxc, etc.); no modification is needed. My second modification occurs towards the end of the UKB file, where the labelled dataset "`labelled_zbnzxi`" is saved:

```
116 data savedata.labelled_zbnzxi; /*LC: I specfied my library savedata here*/
117    set savedata.raw_zbnzxi;  /*LC: I specfied my library savedata here*/
118    format n_50_0_0 13.1 n_50_1_0 13.1 n_50_2_0 13.1 n_52_0_0 miaa. n_189_0_0 22.18
119    n_190_0_0 mpxc. s_191_0_0 date9. n_864_0_0 mzjsf. n_864_1_0 mzjsf. n_864_2_0 mzjsf.
```

I recommend using a log file. If using such a log file, it is good practice to close it at the end of the code.

```
284  * LC: Save the log file, overwriting the previous one;
285  dm 'log; file "K:\TEU\CancerPRS\1stData_12Feb2019\SelectedConvSAS_71fields\
286    ukb71fields_LC_&sysdate9..log" replace';
```

The key modification I have made in this step is that I have specified my own SAS library "savedata" for saving the datasets. You do not have to do so, in which case the resulting SAS datasets will be saved in the default SAS "work" library. Right-clicking on the "work" library shows its location as:

C:\Users\leic\AppData\Local\Temp\SAS Temporary Files_TD1908_NDPH6985_

(NB The "AppData" folder is a hidden folder in Windows.) You will need to go to the folder above to find your datasets, and then copy them to your own folder.

When executed, the "ukb26505_LC.sas" script produces the following three files, of which the first two will be used in your subsequent SAS code.

1. formats.sas7bcat: this is the SAS catalogue (37 KB), needed by your future SAS code for defining the format of the data.
2. labelled_ zbnzxi.sas7bdat: this is the labelled SAS data set (1 GB, 71 fields, 254 variables). This is your dataset to use.
3. raw_zbnzxi.sas7bdat: this is the raw (unlabelled) SAS data set, for record-keeping purposes.

The exemplar code above takes 30 seconds to run when presented with our small example dataset of 71 fields (254 variables). A large dataset of 0.5 million observations with 7000 variables takes about 20 minutes to run.

3. Epilogue

3.1 Choice of statistical language

The main reason for my choice of SAS here is that I would like to rename and group all my variables later using a SAS macro that other researchers in NDPH has developed. The renamed variables can then be imported into other languages (eg R or Stata) if required.

3.2 Using the “ukbgene” and “gtool“ commands

As noted above, the “ukbgene” command (one of the six downloaded file handles) needs to be run in Linux instead of Windows. Our group (TEU) has used “ukbgene” and “gtool” via an Ubuntu Virtual Machine (VM) installed on our Windows PC. Another step-by-step guide “UKB_DataGuide_SNPs” is planned on this subject.