PDS4 | 2018-10-R1

Periodic Dataset 4

Importing Enroll-HD PDS Files

Enroll-HD

A worldwide observational study for Huntington's disease families

A CHDI Foundation Project

Contents

1. PURPOSE OF DOCUMENT

This document provides instructions on how to import the .csv (comma-separated values) formatted Enroll-HD PDS data into Excel and R (a software environment for statistical computing and graphics). The document contains step by step instructions on how to open and format the data files to make the data easy to examine and use. There are other methods that may work equally as well. This document is not intended to be an exhaustive, but to simply provide one method.

2. DATA FILES PROVIDED ENROLL-HD

The Enroll-HD PDS dataset is provided in two formats:

- CSV file: CSV stands for commaseparated values (.csv) which is a delimiter-separated format. There are many types of delimiters including commas, semicolons, tabs, etc. The PDS data uses the tab as the delimiter (⟶ tab).
- R file: binary code format for the R software application.

The .csv file format can be imported into Excel spreadsheets as well as into most statistical software packages including R, Stata, and SAS.

In many cases it is essential to specify, **before importing the data into the analysis software, that the variables are separated by tabs**, as the default delimiter may change based on the country or region where the software was developed. It is also **important that these files should not be edited in a word processing software or other programs that may potentially modify the tab characters, as this may damage the integrity of the original files**. CSV files can be saved in other formats which are compatible with other statistical software packages as needed.

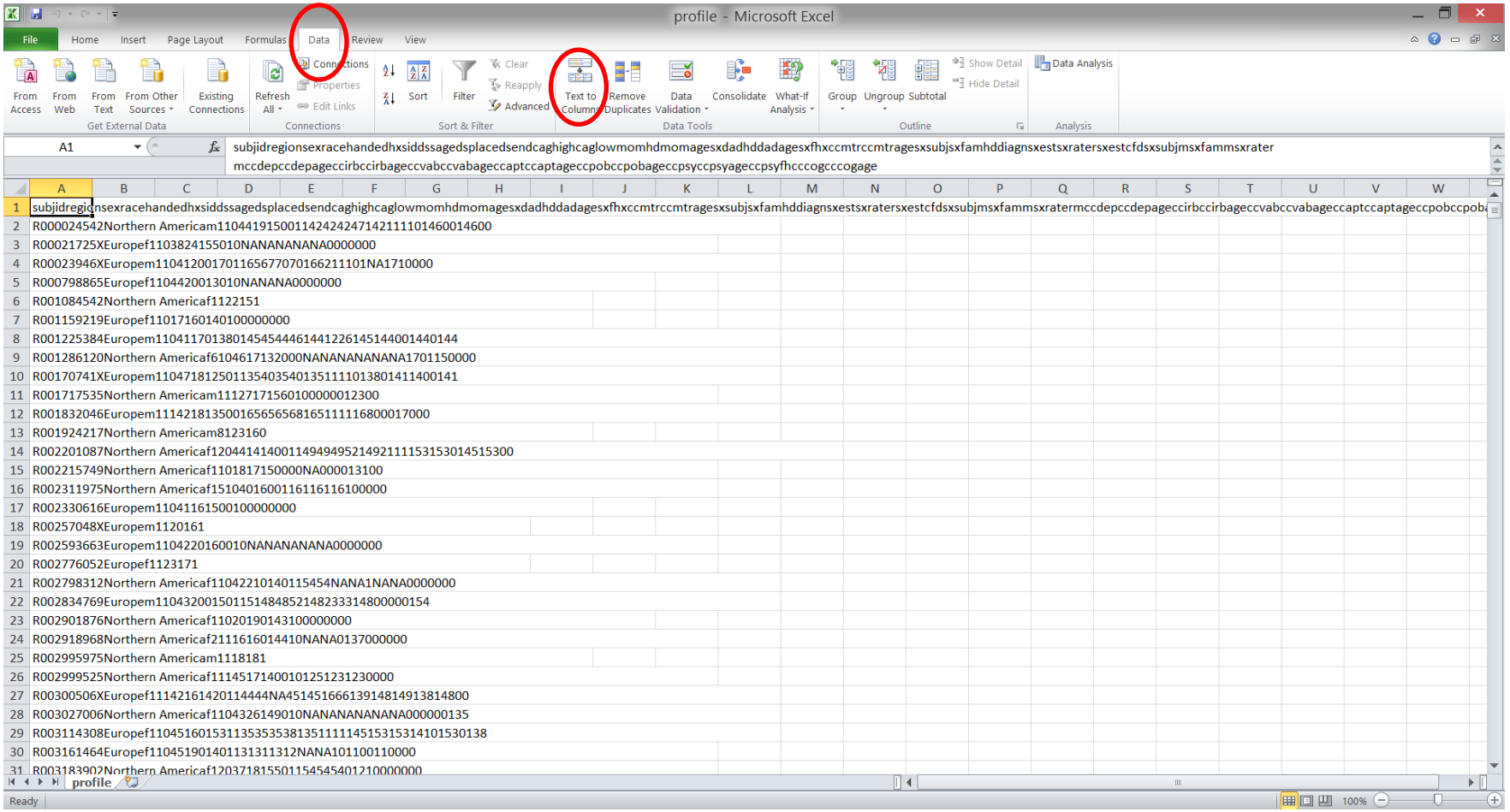
3. IMPORT .csv FILES INTO EXCEL

The .csv files can be easily imported and opened in Microsoft Excel. Because Excel is language dependent and delimiters differ from one country to another, some considerations need to be addressed when opening the .csv files to maintain data integrity. The procedures outlined here, to open the .csv files, can be applied to different versions of Excel.

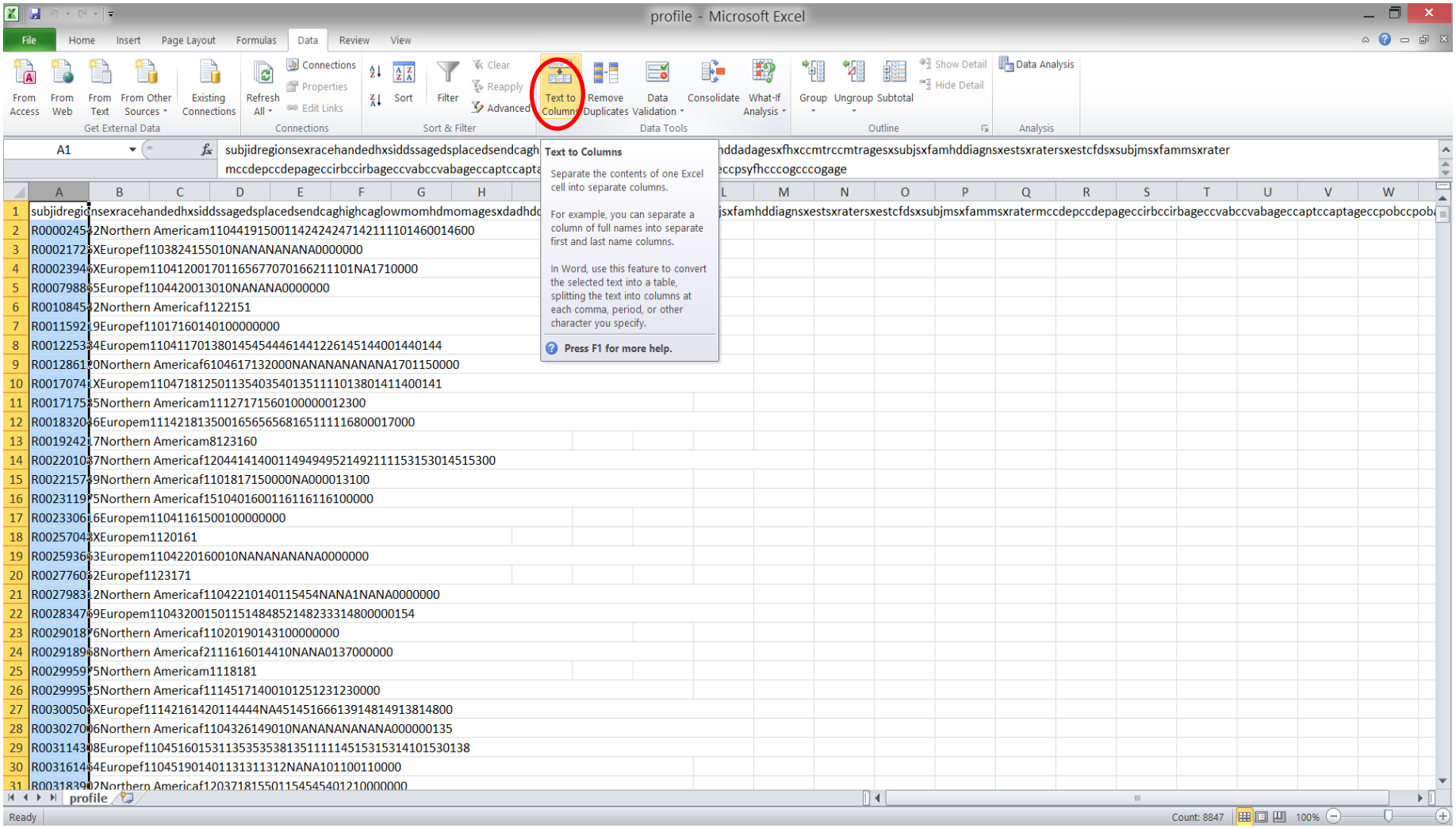
As a default, Excel reads the values for each column as being in a "General" format. For example, unless otherwise specified, Excel interprets figures as numbers (e.g 1234), entered dates as date format (as pre-set, e.g. 11/28/2016), and it changes other values to text format (e.g. Aspirin). For some entries this is counterproductive, as Excel may misinterpret the entries and reformat the data, effectively changing the data (e.g. 1.5 is read as May 1 instead of 1.5 mg; or the WHO-DD Code for Tetrabenazine 00222101003 is changed to 22211003, removing the important leading "0"s). **To maintain the integrity of the data, each data column needs to be carefully examined prior to importing the data into Excel.**

Below are step by step guidelines for correctly importing the data into Excel:

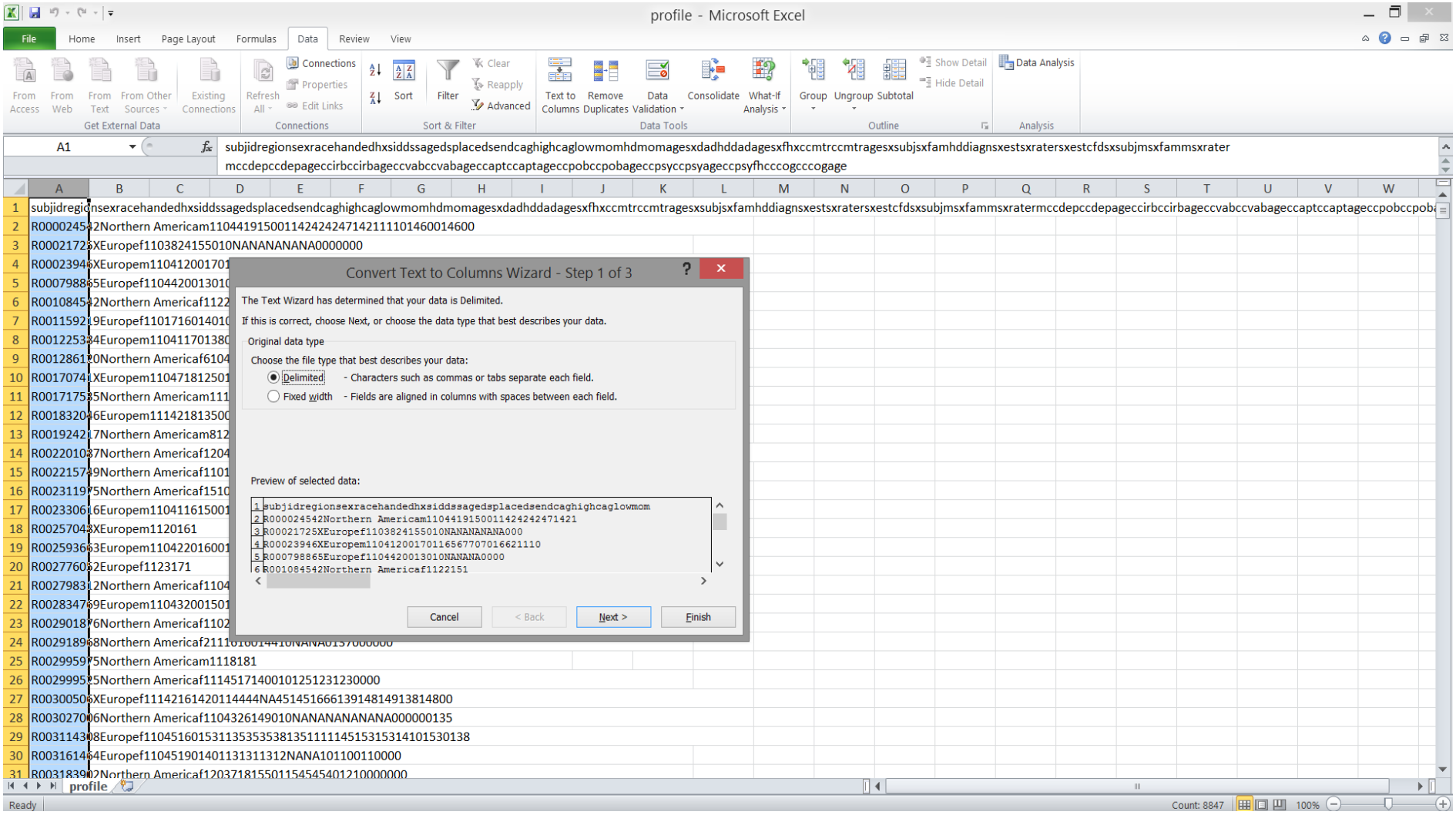
Step 1 – Importing data: Either open .csv file using Excel, or open Excel then go to tab 'Data' and click on get external data 'From Text'. The file used for this demonstration is the 'profile.csv' file. In the first step, when the data is imported all data will be entered into the first column of the file.



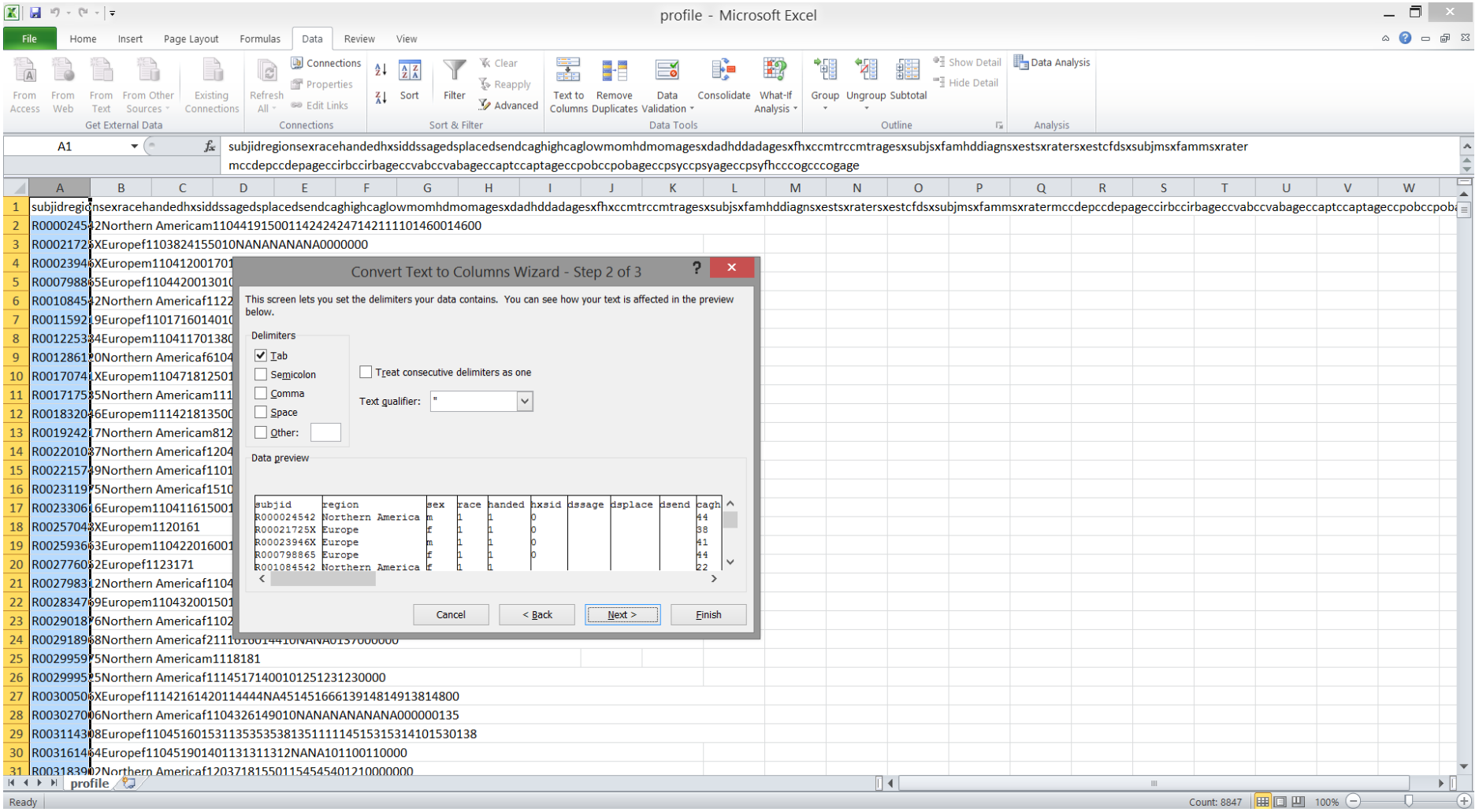
Step 2 – Separating data: Select the first column and on the tab 'Data' click on 'text to column'. A block screen – "wizard" will appear. This wizard will help guide you to proper separation of the data.



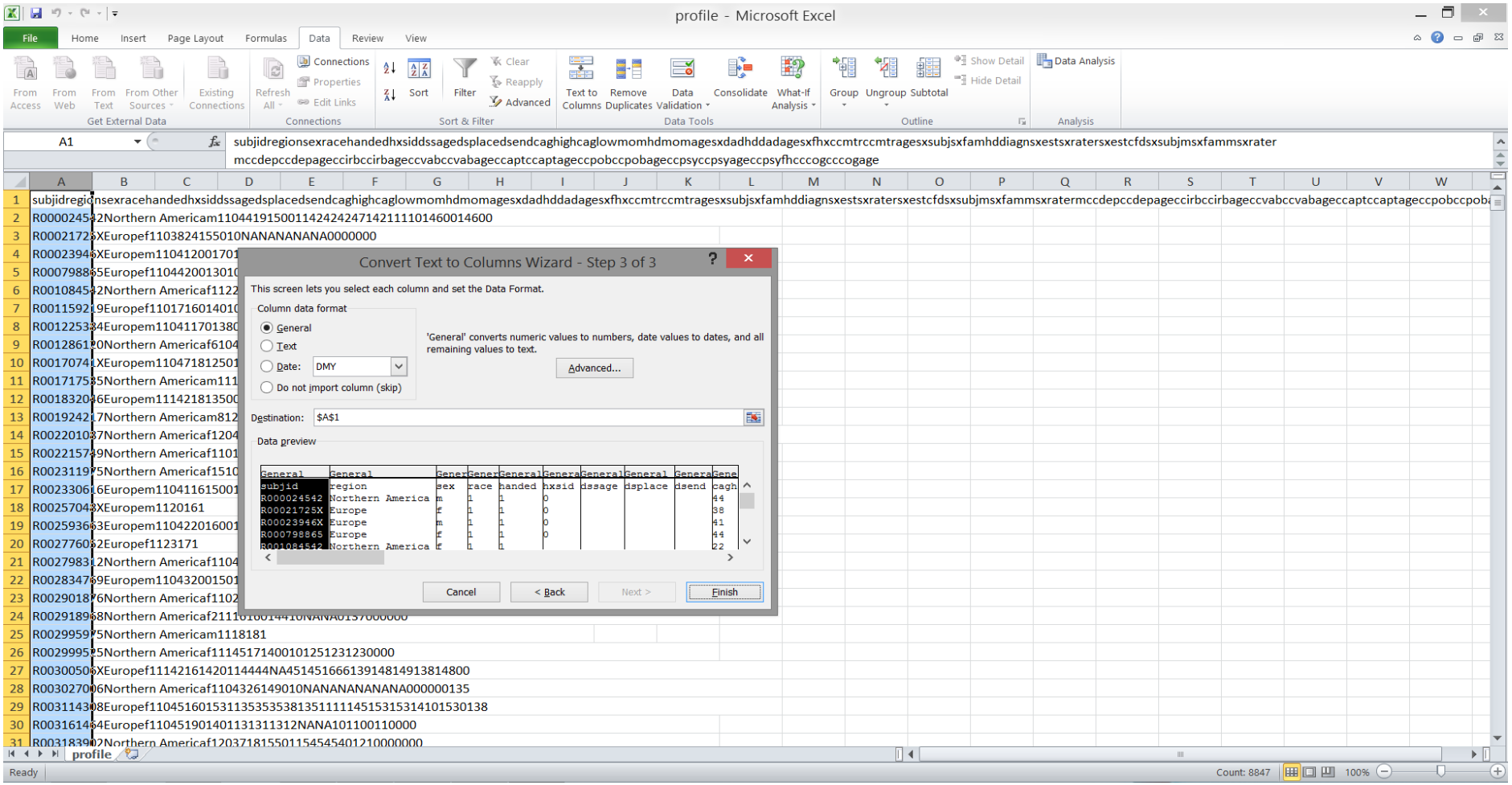
Step 3 – Choose way of separating the data: In the 'text to column' wizard, select the 'delimited' option; this will separate the variables in separate columns and then click 'Next'.



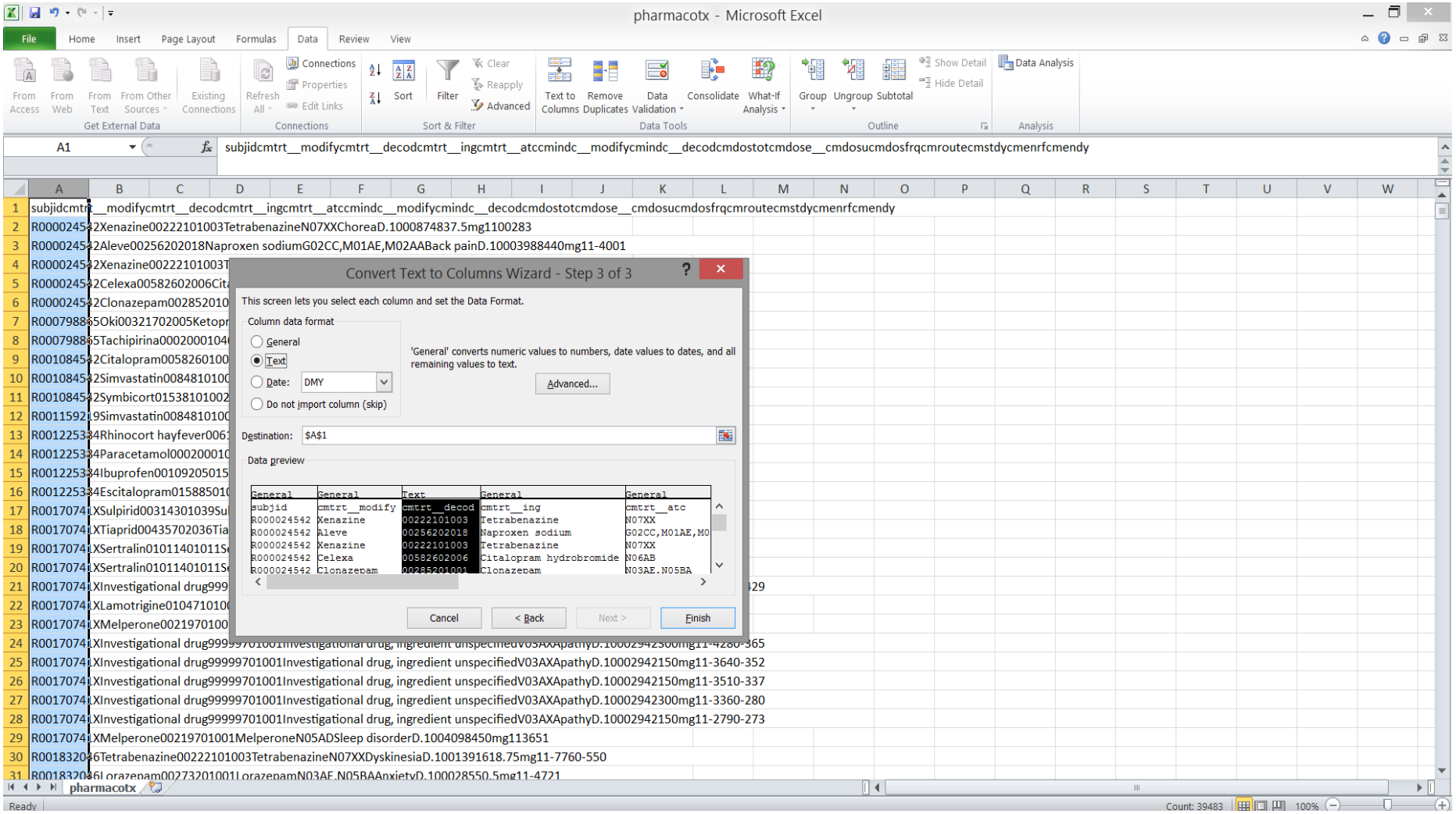
Step 4 – Choose delimiter: Selecting the "Delimited" radio button lets Excel know that the data is separated by commas or tabs. Once you click next the wizard will provide a preview of how the data will be separated into different columns..Then click "Next" to proceed.



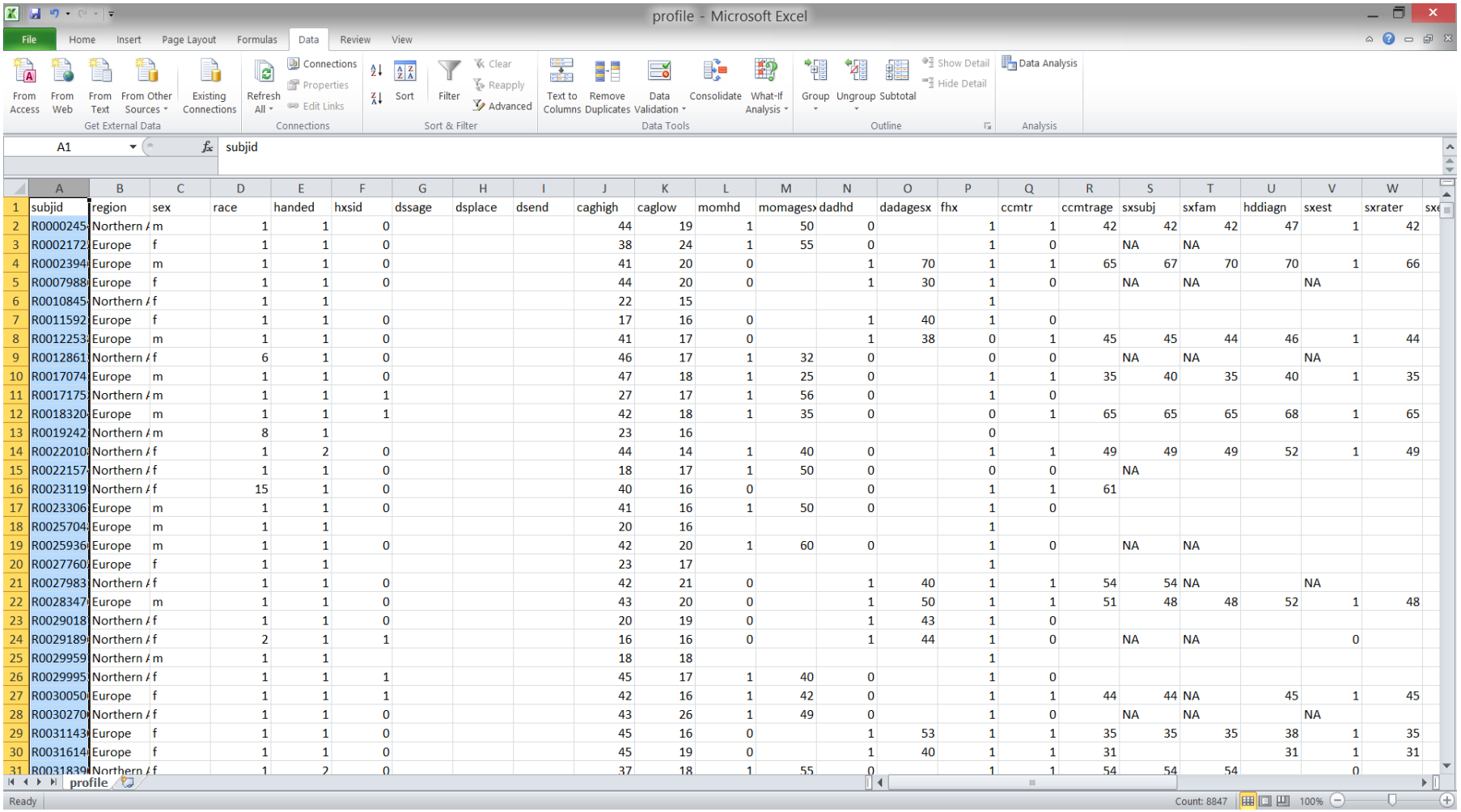
Step 5 – Assign formats: For each column a format needs to be assigned. The default format 'General' works for most columns. Columns where numbers have leading "0" and columns with mixed entries like 1.5, 1,5, 1/5, need to be explicitly formatted as Text, as entries might otherwise become corrupted in an unchangeable way. After assigning the format to each column click 'Finish'.



NOTE: The data files pharmacotx and nutsuppl contain two columns ‘cmtrt_decod’ and ‘cmdostot’ that require formatting as Text.



Step 6: The .csv file is column-separated and should be saved as an Excel file (.xls or .xlsx) using the 'Save As' option.



subjid	region	sex	race	handed	hxsid	dssage	dsplace	dsend	caghigh	caglow	momhd	momages	dadhd	dadagesx	fhx	ccmtr	ccmtrage	sxsubj	sxfam	hddiagn	sxest	sxrater
R0000245	Northern A	m	1	1	0				44	19	1	50	0		1	1	42	42	42	47	1	42
R0002172	Europe	f	1	1	0				38	24	1	55	0		1	0		NA	NA			
R0002394	Europe	m	1	1	0				41	20	0		1	70	1	1	65	67	70	70	1	66
R0007988	Europe	f	1	1	0				44	20	0		1	30	1	0		NA	NA		NA	
R0010845	Northern A	f	1	1					22	15					1							
R0011592	Europe	f	1	1	0				17	16	0		1	40	1	0						
R0012253	Europe	m	1	1	0				41	17	0		1	38	0	1	45	45	44	46	1	44
R0012861	Northern A	f	6	1	0				46	17	1	32	0		0	0		NA	NA		NA	
R0017074	Europe	m	1	1	0				47	18	1	25	0		1	1	35	40	35	40	1	35
R0017175	Northern A	m	1	1	1				27	17	1	56	0		1	0						
R0018320	Europe	m	1	1	1				42	18	1	35	0		0	1	65	65	65	68	1	65
R0019242	Northern A	m	8	1					23	16					0							
R0022010	Northern A	f	1	2	0				44	14	1	40	0		1	1	49	49	49	52	1	49
R0022157	Northern A	f	1	1	0				18	17	1	50	0		0	0		NA				
R0023119	Northern A	f	15	1	0				40	16	0		0		1	1	61					
R0023306	Europe	m	1	1	0				41	16	1	50	0		1	0						
R0025704	Europe	m	1	1					20	16					1							
R0025936	Europe	m	1	1	0				42	20	1	60	0		1	0		NA	NA			
R0027760	Europe	f	1	1					23	17					1							
R0027983	Northern A	f	1	1	0				42	21	0		1	40	1	1	54	54	NA		NA	
R0028347	Europe	m	1	1	0				43	20	0		1	50	1	1	51	48	48	52	1	48
R0029018	Northern A	f	1	1	0				20	19	0		1	43	1	0						
R0029189	Northern A	f	2	1	1				16	16	0		1	44	1	0		NA	NA		0	
R0029959	Northern A	m	1	1					18	18					1							
R0029995	Northern A	f	1	1	1				45	17	1	40	0		1	0						
R0030050	Europe	f	1	1	1				42	16	1	42	0		1	1	44	44	NA	45	1	45
R0030270	Northern A	f	1	1	0				43	26	1	49	0		1	0		NA	NA		NA	
R0031143	Europe	f	1	1	0				45	16	0		1	53	1	1	35	35	35	38	1	35
R0031614	Europe	f	1	1	0				45	19	0		1	40	1	1	31			31	1	31
R0031839	Northern A	f	1	2	0				37	18	1	55	0		1	1	54	54	54		0	

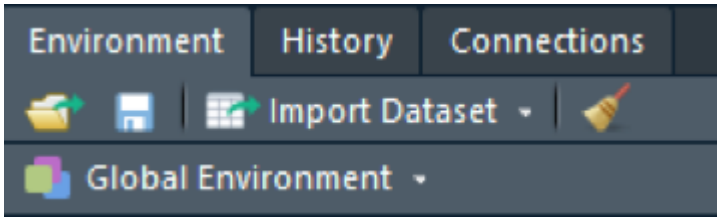
4. IMPORT .csv INTO R

First, a package capable of reading CSV files has to be loaded into R environment. For example "readr" is one of the most popular applications. If a package like “readr” is not already installed, the CVS data files can be imported using the following code line: **`install.packages(readr)`**. Then, to load the CSV data into R using a package like “readr” use the: **`library(readr)`** command. To ensure the CSV file is imported correctly, set the directory to the file folder where the PDS files are located, and then run the following code:

```
file = read_delim("file.csv", "\t", escape_double = FALSE,
trim_ws = TRUE)
```

5. IMPORT R data FILES INTO R

This data file is specific for R. After loading the R data files into R, 9 data frames are made available in the R environment and are ready to be used. The loading can be done using the function **`load(“Rdata_directory”)`**. For Rstudio users, the loading can be performed by clicking in the “load workspace” ribbon, and then browsing for the location of the R data file.



Revision History

Document Name	Summary of Changes
Version 2015-10-R1	Initial version for second Enroll-HD periodic dataset
Version 2016-10-R1	Revised version for third Enroll-HD periodic dataset
Version 2018-10-R1	Third version for fourth Enroll-HD periodic dataset