1. File formats

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

- **CSV file:** CSV stands for comma separated values (.csv) which is a delimiter-separated format. The PDS data uses the **tab** as the delimiter. Software settings need to be adapted respectively.
- **R file**: binary code format for the R[1] software application (a software environment for statistical analysis).

Because of the complexity and the size of the data set, use of a statistical software package such as R, Stata, or SAS is recommended. The .csv file format can also be imported into Excel (caution is advisable).

It is important that files **are not be edited in a word processing software or other programs that may potentially modify 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.

[1] R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

2. Importing data

Importing CSV files into Excel

The .csv files can be 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 most recent 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 numeric data as numbers (e.g., 1234), entered dates as date format (as pre-set, e.g. 11/28/2016), and changes other values (e.g. strings) to text format (e.g. Aspirin). For some entries this is counterproductive, as Excel may **misinterpret entries** and **incorrectly 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**.

An illustrated guide for correctly importing CSV data files into Excel are provided in **Appendix A.**

Importing CSV files into R

Make sure the CSV file has not been opened and saved using a word processing software. A software package capable of reading CSV files must be loaded into R environment. The package "readr" is one of the most popular packages, but there are several others that will also work. If a package like "readr" is not already installed, the CSV data files can be imported using the following code line:

```
install.packages(readr)
```

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)
```

Importing R 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 command:

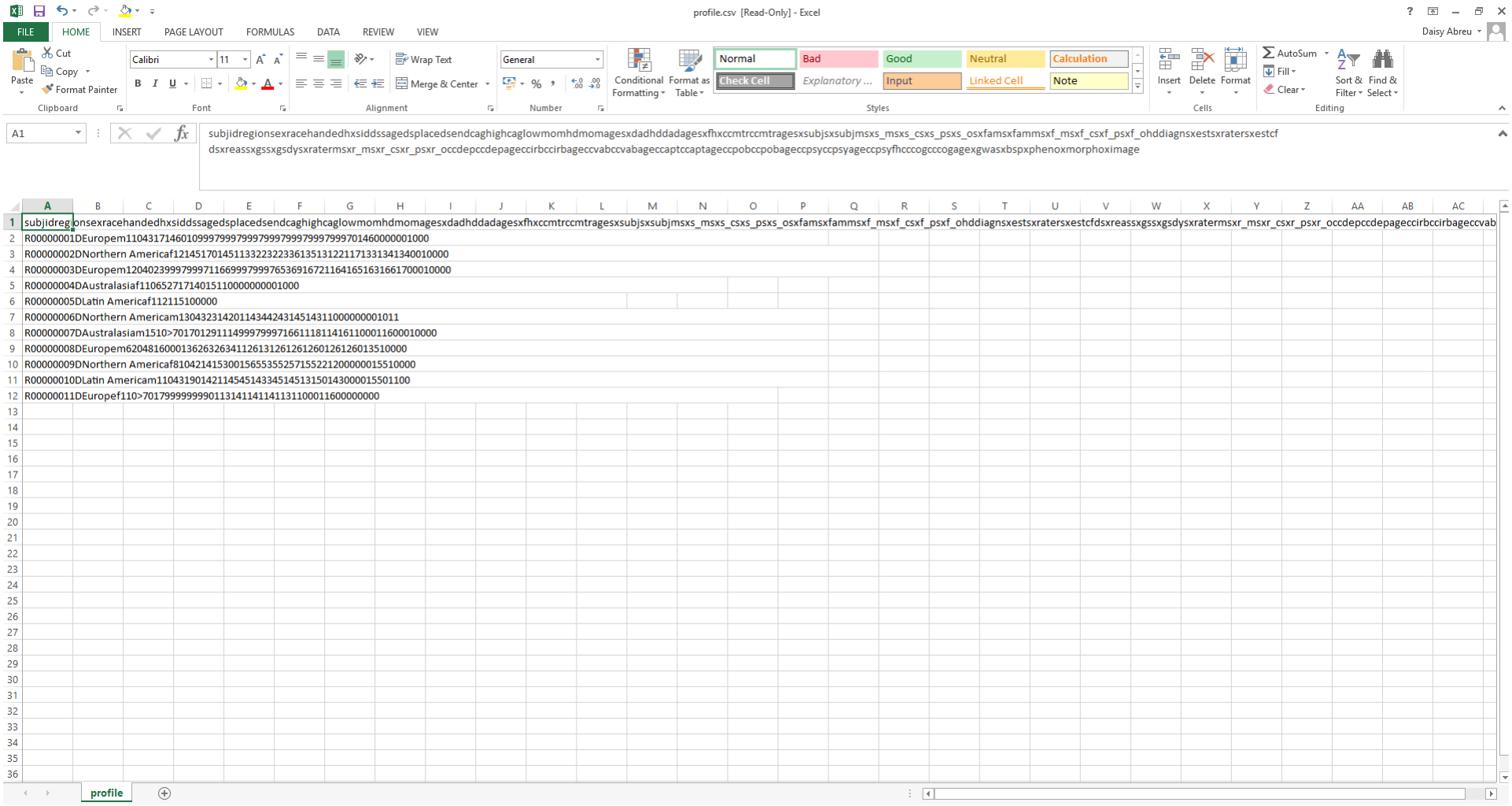
```
load("Rdata_directory")
```

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.

Appendix A: An illustrated guide to correctly importing CSV files into Excel

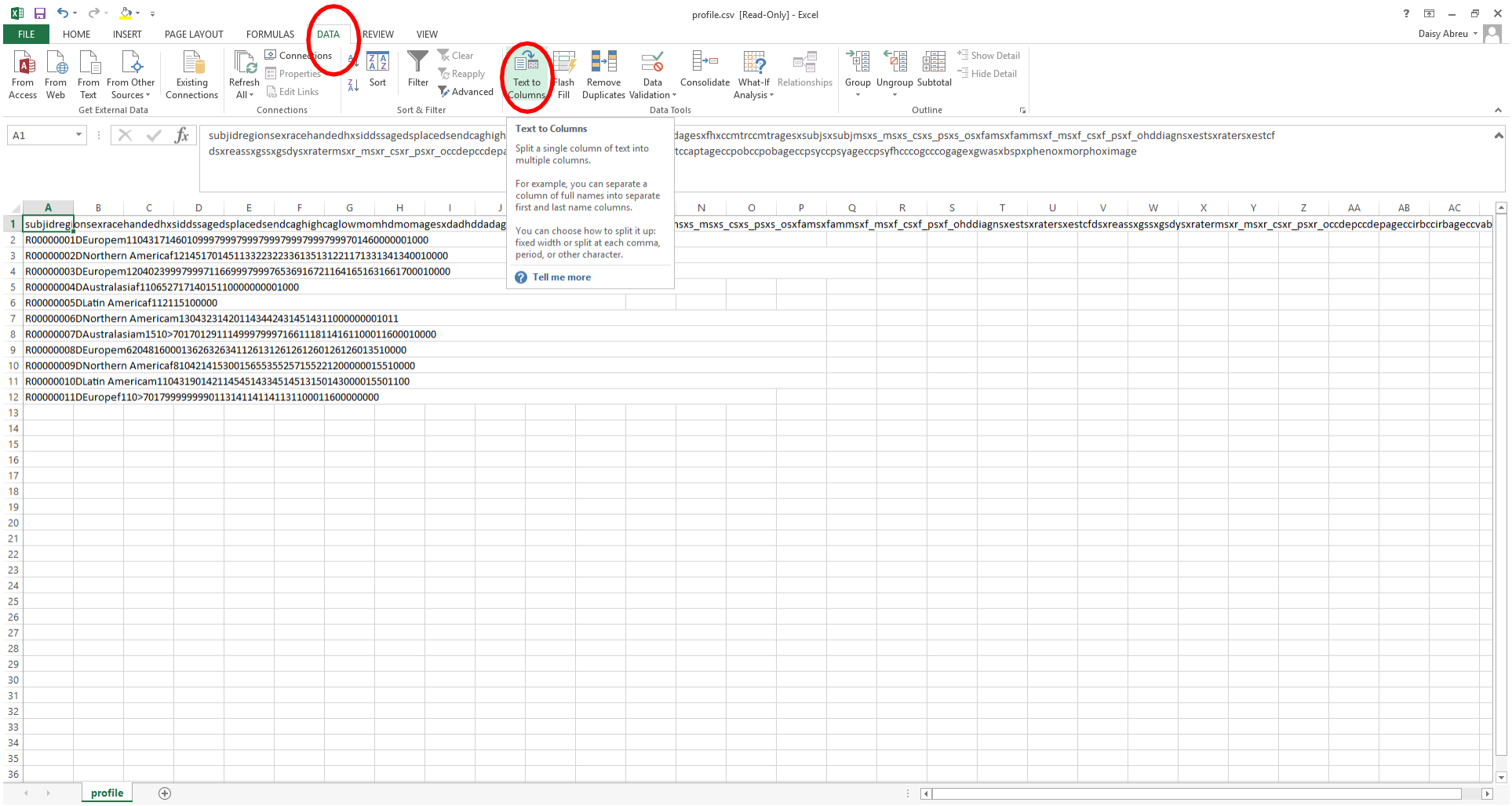
The file used for this demonstration is the 'profile.csv' file.

Step 1 – Open CSV file in Excel: Open the .csv file using Excel, or open Excel and on the "Data" tab click "**From Text/CSV**". Data will be imported in entirety into the first column of the Excel file, as illustrated below.

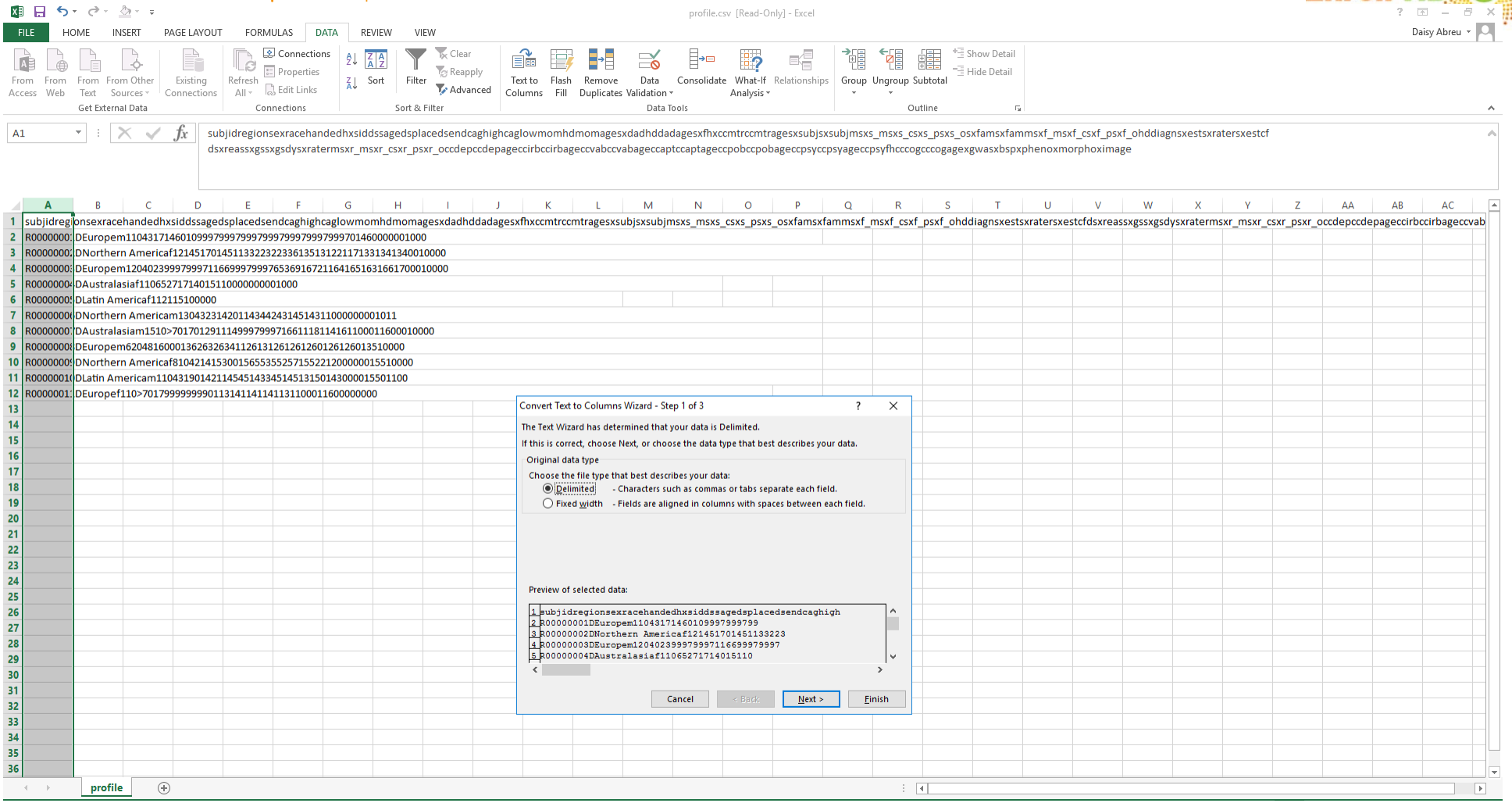


Step 2 – Open Text to Columns Wizard: Select the first column, then on the tab "Data" click "**Text to Columns**". A wizard will appear to guide you through the

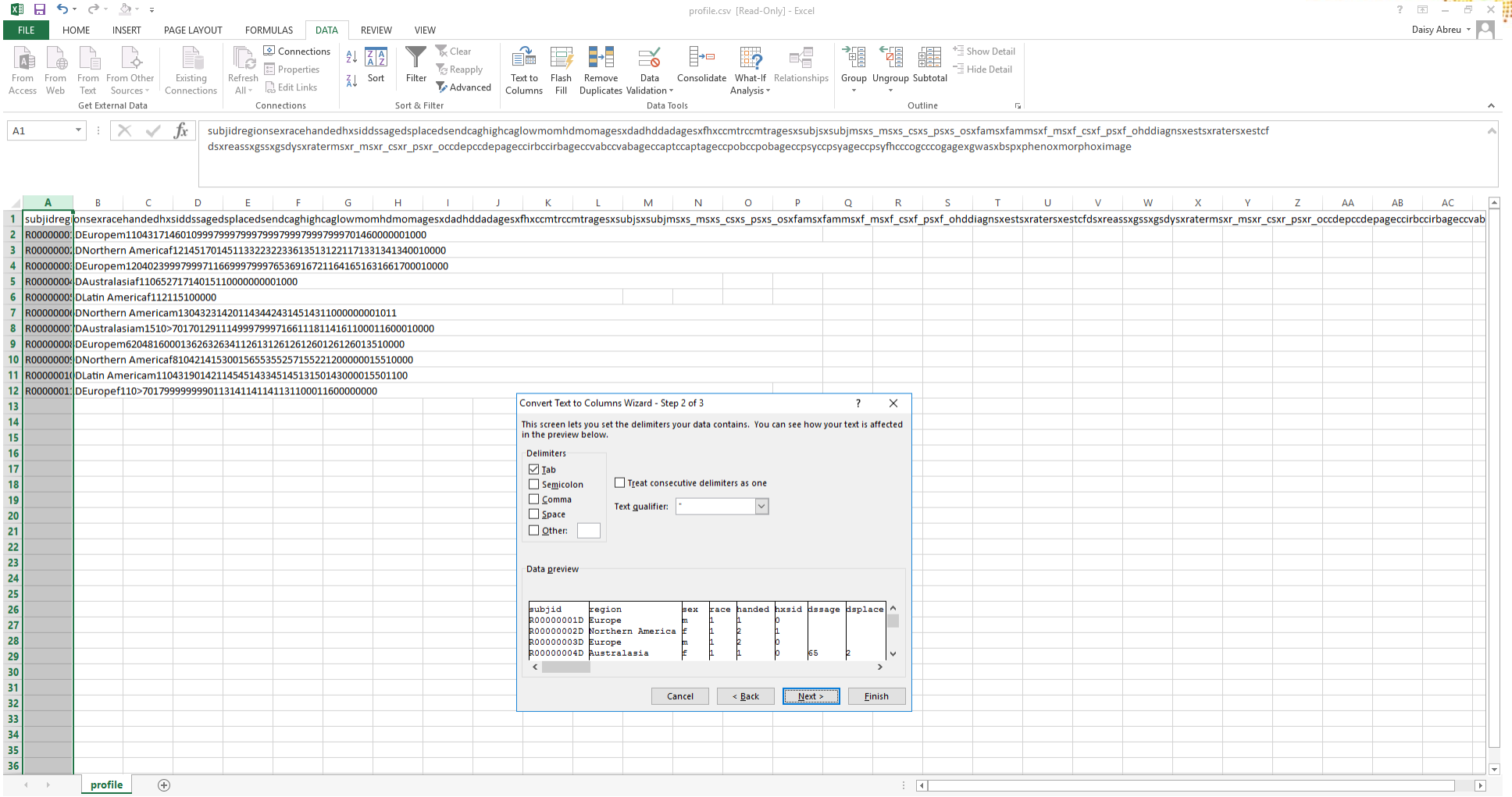
process.



Step 3 – Specify data file type: In the Text to Columns Wizard (step 1 of 3), select the "**Delimited**" checkbox (this lets Excel know that the data fields are separated by commas or tabs), then click "**Next**".

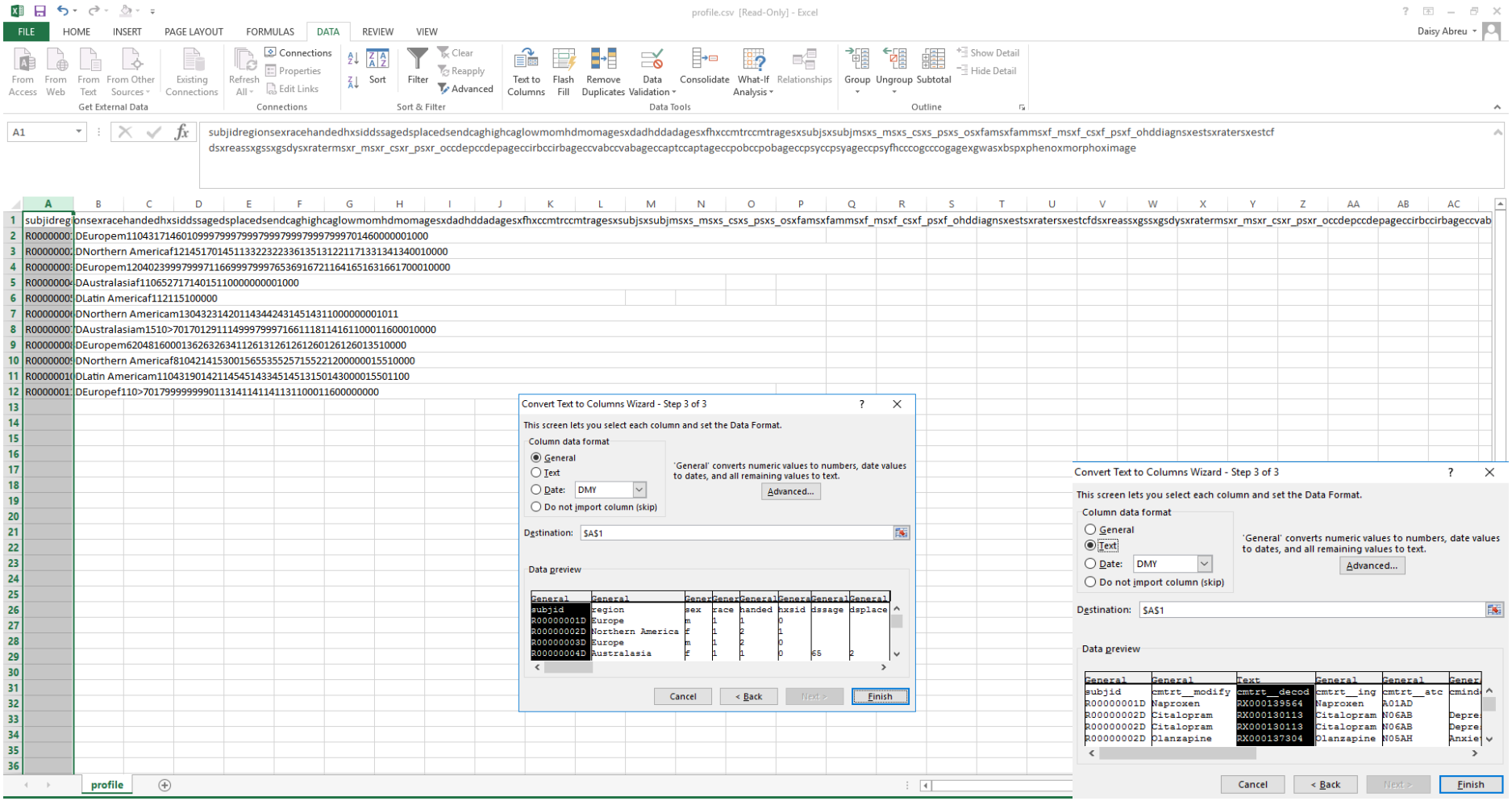


Step 4 – Select delimiter type: In the Text to Columns Wizard (step 2 of 3), select the Delimiter type "**Tab**" (this lets Excel know that the data fields are separated by tabs specifically), then click "**Next**".



Step 5 – Assign column formats: For each column (i.e., variable), an appropriate format needs to be assigned. This is completed in the Text to Columns Wizard (step 3 of 3). 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,

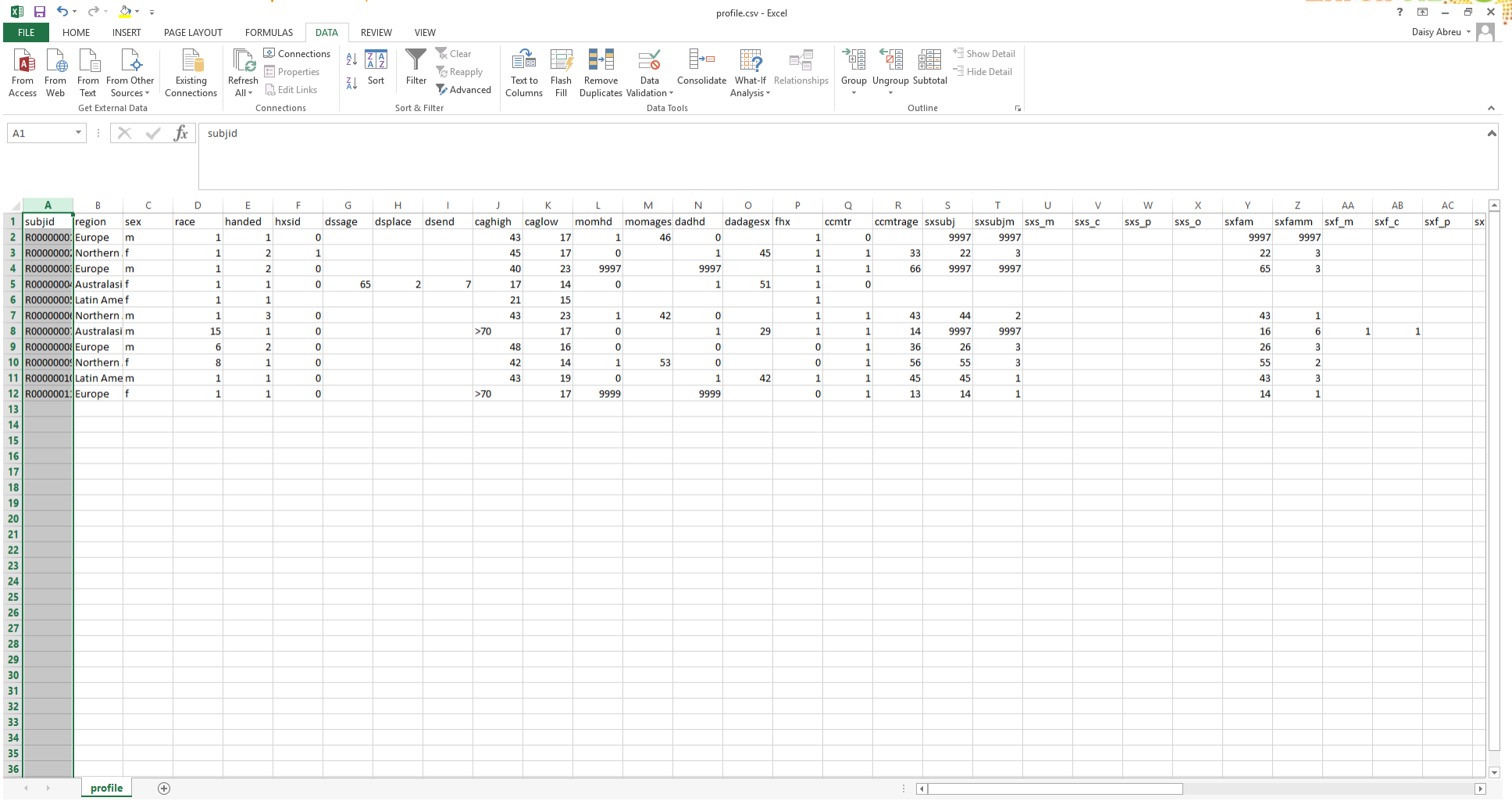
need to be explicitly formatted as "**Text**", as entries might otherwise become corrupted in an unchangeable way. After assigning the correct format to each column, click "**Finish**".



NB: The data files *pharmacotx* and *nutsuppl* contain two columns '*cmtrt_decod*' and '*cmdostot*' that require formatting as "**Text**".

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

Enroll-HD: Download and Import Data | 2022-05-25



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
1	subjid	region	sex	race	handed	hxsid	dssage	dsplace	dsend	caghigh	caglow	momhd	momages	dadhd	dadagesx	fhx	ccmtr	ccmtrage	sxsubj	sxsubjm	sxs_m	sxs_c	sxs_p	sxs_o	sxfam	sxfamm	sxf_m	sxf_c	sxf_p	sx
2	R0000000:	Europe	m	1	1	0				43	17	1	46	0		1	0		9997	9997					9997	9997				
3	R0000000:	Northern	f	1	2	1				45	17	0		1	45	1	1	33	22	3					22	3				
4	R0000000:	Europe	m	1	2	0				40	23	9997		9997		1	1	66	9997	9997					65	3				
5	R0000000	Australasi	f	1	1	0	65	2	7	17	14	0		1	51	1	0													
6	R0000000:	Latin Ame	f	1	1					21	15					1														
7	R0000000	Northern	m	1	3	0				43	23	1	42	0		1	1	43	44	2					43	1				
8	R0000000:	Australasi	m	15	1	0				>70	17	0		1	29	1	1	14	9997	9997					16	6	1	1		
9	R0000000	Europe	m	6	2	0				48	16	0		0		0	1	36	26	3					26	3				
10	R0000000	Northern	f	8	1	0				42	14	1	53	0		0	1	56	55	3					55	2				
11	R0000001	Latin Ame	m	1	1	0				43	19	0		1	42	1	1	45	45	1					43	3				
12	R0000001:	Europe	f	1	1	0				>70	17	9999		9999		0	1	13	14	1					14	1				
13																														
14																														
15																														
16																														
17																														
18																														
19																														
20																														
21																														
22																														
23																														
24																														
25																														
26																														
27																														
28																														
29																														
30																														
31																														
32																														
33																														
34																														
35																														
36																														