



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.....	3
2. DATA FILES PROVIDED ENROLL-HD.....	3
3. IMPORT .csv FILES INTO EXCEL	3
4. IMPORT .csv INTO R.....	12
5. IMPORT Rdata FILES INTO R.....	12

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.

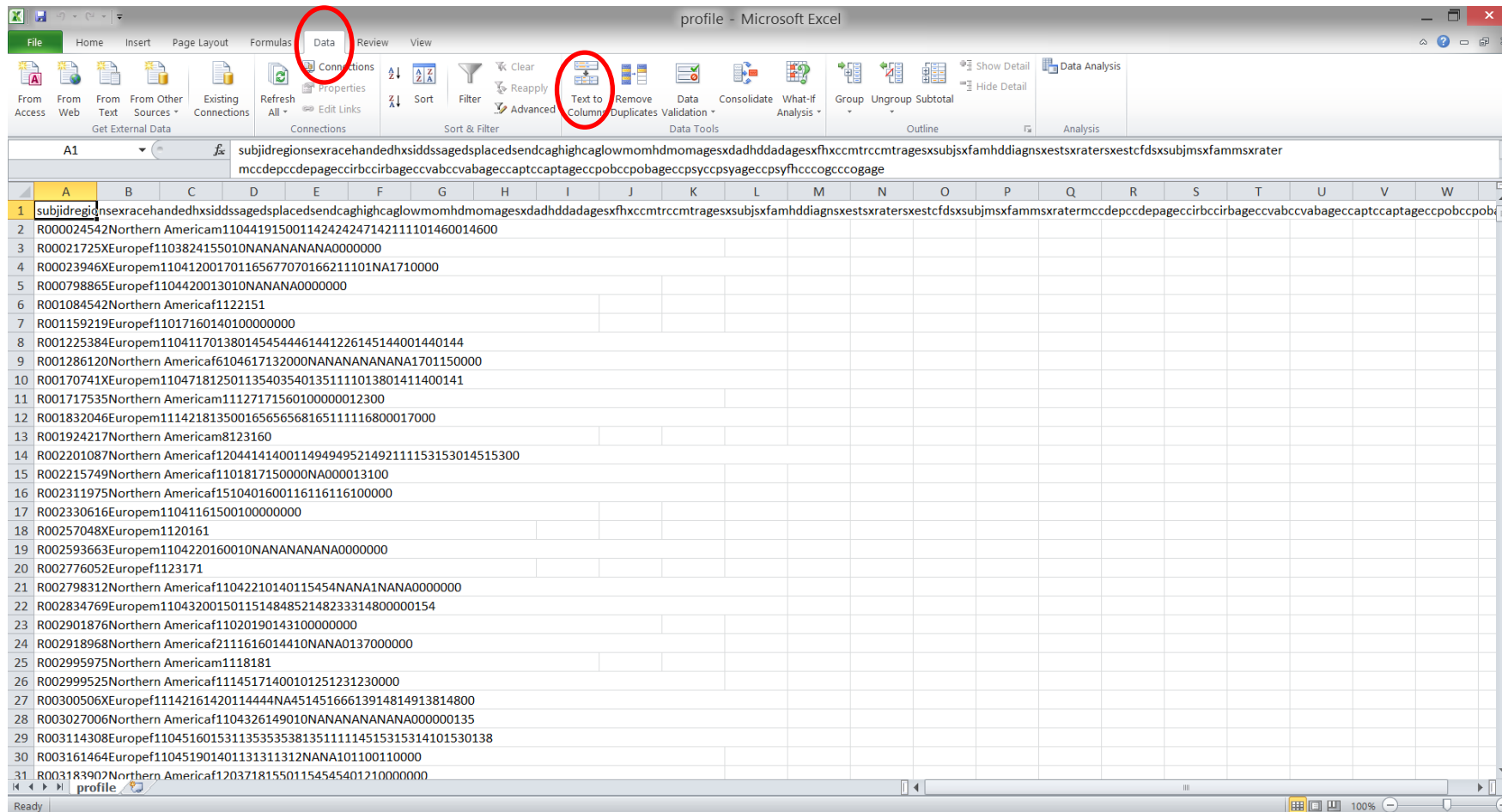
3. IMPORT .csv FILES INTO EXCEL

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.

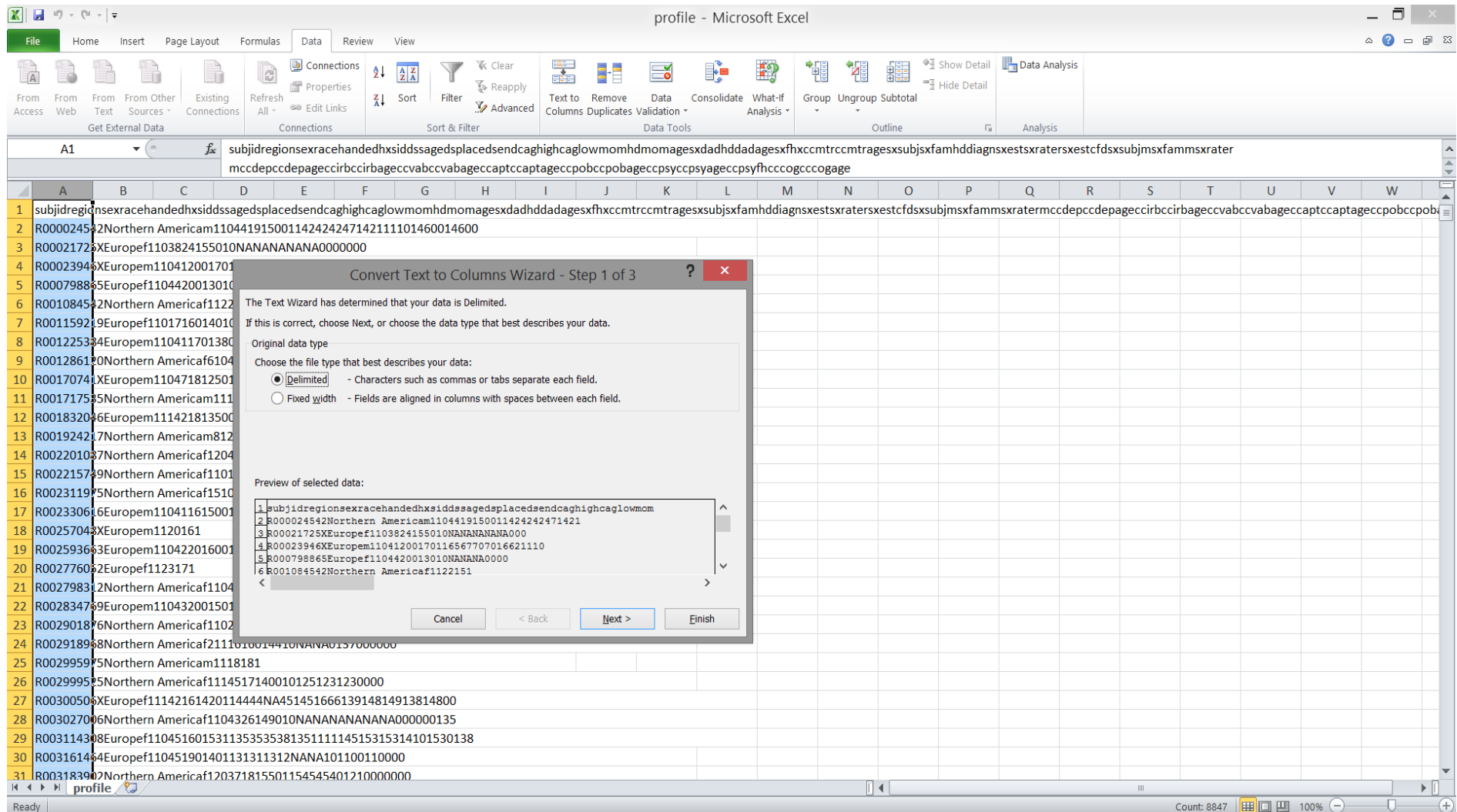


Step 2 – Separating data: Select the first column and on the tab ‘Data’ click on ‘text to column’. A block screen – “wizard” will appear. This wizard will help guide you to proper separation of the data.

The screenshot shows the Microsoft Excel interface with the 'Data' tab selected. The 'Text to Columns' button in the 'Data Tools' group is highlighted with a red circle. The 'Text to Columns' wizard is open, displaying the 'Text to Columns' dialog box. The data in the spreadsheet is a list of IDs and names, such as 'R00002452 Northern American' and 'R00021723 Europe'.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	subjidregionsexracehandedhxssiddssagedsplacedsendcagh																						
2	R00002452	Northern American	110441915001142424247142111101460014600																				
3	R00021723	Europe	1103824155010NANANANANA0000000																				
4	R00023943	Europe	110412001701165677070166211101NA1710000																				
5	R00079885	Europe	1104420013010NANANANA0000000																				
6	R00108454	Northern American	1122151																				
7	R00115921	Europe	11017160140100000000																				
8	R00122533	Europe	1104117013801454544461441226145144001440144																				
9	R00128612	Northern American	1104617132000NANANANANA1701150000																				
10	R00170741	Europe	1104718125011354035401351111013801411400141																				
11	R00171753	Northern American	11127171560100000012300																				
12	R00183204	Europe	111421813500165656816511116800017000																				
13	R00192421	Northern American	8123160																				
14	R00220103	Northern American	1204414140011494949521492111153153014515300																				
15	R00221579	Northern American	1101817150000NA000013100																				
16	R00231195	Northern American	1510401600116116116100000																				
17	R00233061	Europe	11041161500100000000																				
18	R00257043	Europe	1120161																				
19	R00259363	Europe	1104220160010NANANANANA0000000																				
20	R00277603	Europe	1123171																				
21	R00279831	Northern American	11042210140115454NANA1NANA0000000																				
22	R00283479	Europe	110432001501151484852148233314800000154																				
23	R00290187	Northern American	11020190143100000000																				
24	R00291898	Northern American	2111616014410NANA0137000000																				
25	R00299597	Northern American	1118181																				
26	R00299955	Northern American	11145171400101251231230000																				
27	R00300506	Europe	11142161420114444NA45145166613914814913814800																				
28	R00302706	Northern American	1104326149010NANANANANANA000000135																				
29	R00311430	Europe	11045160153113535353813511114515315314101530138																				
30	R00316144	Europe	11045190140113131312NANA101100110000																				
31	R00318390	Northern American	120371815501154545401210000000																				

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.

The screenshot shows the Microsoft Excel interface with the 'Convert Text to Columns Wizard - Step 2 of 3' dialog box open. The dialog box has a 'Delimiters' section with the following options:

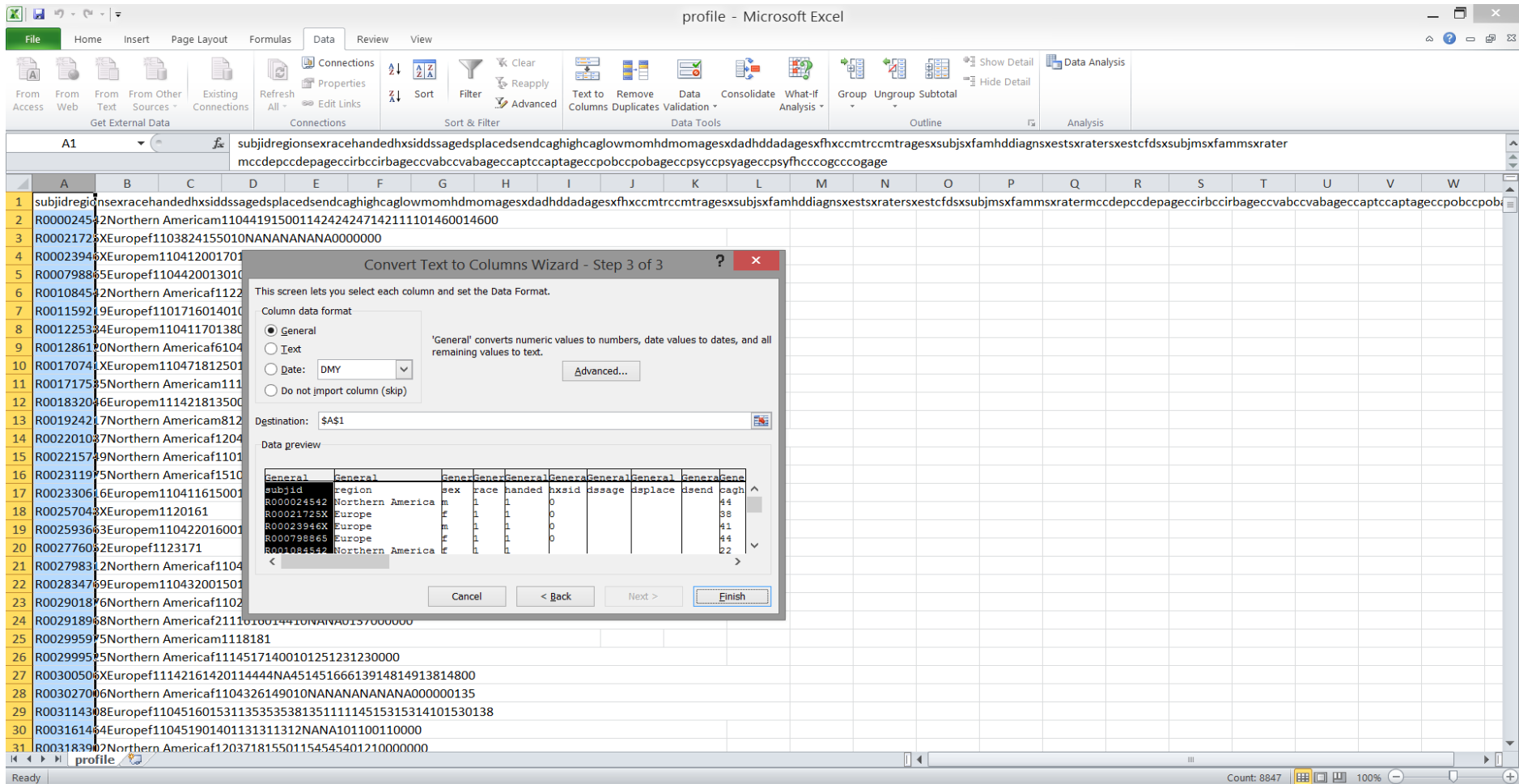
- ☒ Tab
- ☐ Semicolon
- ☐ Comma
- ☐ Space
- ☐ Other: []

There is also a checkbox for 'Treat consecutive delimiters as one' which is unchecked. A 'Text qualifier' dropdown is set to '"'. Below the delimiters is a 'Data preview' section showing a table with the following columns: subj, id, region, sex, race, handed, hxsid, hssage, dsplace, dsend, cagh. The preview shows data for rows 17 through 21 of the spreadsheet.

The spreadsheet data visible in the background includes the following rows (from row 17 to row 31):

Row	Cell A1
17	R000024542 Northern America 110411615001
18	R00021725X Europe 1120161
19	R002593663 Europe 110422016001
20	R000798865 Europe 1123171
21	R001084542 Northern America 110411615001
22	R002798312 Northern America 110411615001
23	R002834789 Europe 110432001501
24	R002901876 Northern America 110411615001
25	R002918968 Northern America 2111010014410NANA0137000000
26	R002995975 Northern America 1118181
27	R002995975 Northern America 11145171400101251231230000
28	R00300506X Europe 11142161420114444NA45145166613914814913814800
29	R003027006 Northern America 1104326149010NANA0000000135
30	R003114308 Europe 110451601531135353538135111114515315314101530138
31	R003161484 Europe 110451901401131311312NANA101100110000

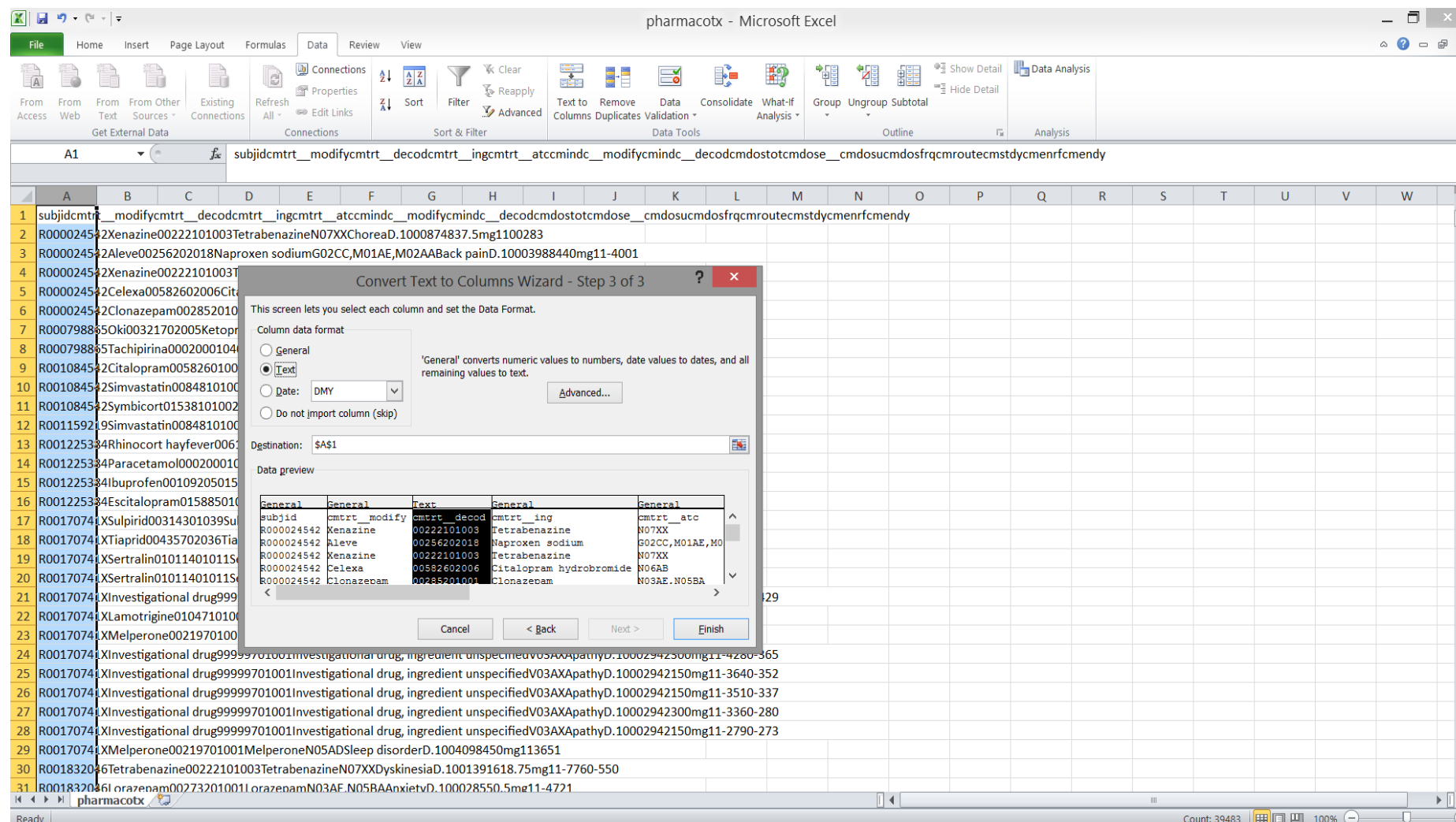
Step 5 – Assign formats: For each column a format needs to be assigned. The default format ‘General’ works for most columns. Columns where numbers have leading “0” and columns with mixed entries like 1.5, 1/5, 1/5, need to be explicitly formatted as Text, as entries might otherwise become corrupted in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



The screenshot shows the 'Convert Text to Columns Wizard - Step 3 of 3' dialog box in Microsoft Excel. The dialog is overlaid on a spreadsheet with data. The 'Column data format' section has 'General' selected. The 'Destination' is '\$A\$1'. The 'Data preview' section shows a table of data with columns labeled 'General' and 'Text'.

General	General	General	General	General	General	General	General	General	General
subjid	region	sex	race	handed	hxsid	dsage	dsplace	dsend	cagh
R000024542	Northern America	m	L	L	0				44
R00021725X	Europe	f	L	L	0				38
R00023946X	Europe	m	L	L	0				41
R000798465	Europe	f	L	L	0				44
R001084542	Northern America	f	L	L	0				22

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.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	
	subjid	region	sex	race	handed	hxsid	dssage	dsplace	dsend	caghigh	caglow	momhd	momages	dadhd	dadagesx	fhx	ccmtr	ccmtrage	sxsubj	sxfam	hddiagn	sxest	srxater	sx
1	R0000245	Northern / m		1	1	0				44	19	1	50	0		1	1	42	42	42	47	1	42	
2	R0002172	Europe	f	1	1	0				38	24	1	55	0		1	0		NA	NA				
3	R0002394	Europe	m	1	1	0				41	20	0		1	70	1	1	65	67	70	70	1	66	
4	R0007988	Europe	f	1	1	0				44	20	0		1	30	1	0		NA	NA		NA		
5	R0010845	Northern / f		1	1					22	15					1								
6	R0011592	Europe	f	1	1	0				17	16	0		1	40	1	0							
7	R0012253	Europe	m	1	1	0				41	17	0		1	38	0	1	45	45	44	46	1	44	
8	R0012861	Northern / f		6	1	0				46	17	1	32	0		0	0		NA	NA		NA		
9	R0017074	Europe	m	1	1	0				47	18	1	25	0		1	1	35	40	35	40	1	35	
10	R0017175	Northern / m		1	1	1				27	17	1	56	0		1	0							
11	R0018320	Europe	m	1	1	1				42	18	1	35	0		0	1	65	65	65	68	1	65	
12	R0019242	Northern / m		8	1					23	16					0								
13	R0022010	Northern / f		1	2	0				44	14	1	40	0		1	1	49	49	49	52	1	49	
14	R0022157	Northern / f		1	1	0				18	17	1	50	0		0	0		NA					
15	R0023119	Northern / f		15	1	0				40	16	0		0		1	1	61						
16	R0023306	Europe	m	1	1	0				41	16	1	50	0		1	0							
17	R0025704	Europe	m	1	1					20	16					1								
18	R0025936	Europe	m	1	1	0				42	20	1	60	0		1	0		NA	NA				
19	R0027760	Europe	f	1	1					23	17					1								
20	R0027983	Northern / f		1	1	0				42	21	0		1	40	1	1	54	54	NA		NA		
21	R0028347	Europe	m	1	1	0				43	20	0		1	50	1	1	51	48	48	52	1	48	
22	R0029018	Northern / f		1	1	0				20	19	0		1	43	1	0							
23	R0029189	Northern / f		2	1	1				16	16	0		1	44	1	0		NA	NA		0		
24	R0029959	Northern / m		1	1					18	18					1								
25	R0029995	Northern / f		1	1	1				45	17	1	40	0		1	0							
26	R0030050	Europe	f	1	1	1				42	16	1	42	0		1	1	44	44	NA		45	1	45
27	R0030270	Northern / f		1	1	0				43	26	1	49	0		1	0		NA	NA		NA		
28	R0031143	Europe	f	1	1	0				45	16	0		1	53	1	1	35	35	35	38	1	35	
29	R0031614	Europe	f	1	1	0				45	19	0		1	40	1	1	31			31	1	31	
30	R0031839	Northern / f		1	2	0				37	18	1	55	0		1	1	54	54	54		0		

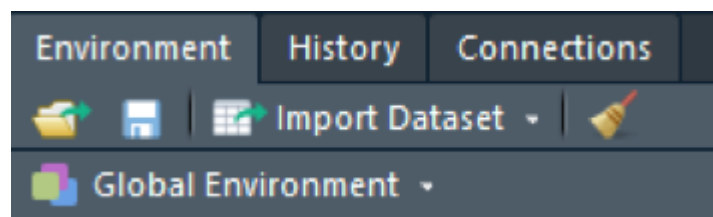
4. IMPORT .csv INTO R

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 CSV 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