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:

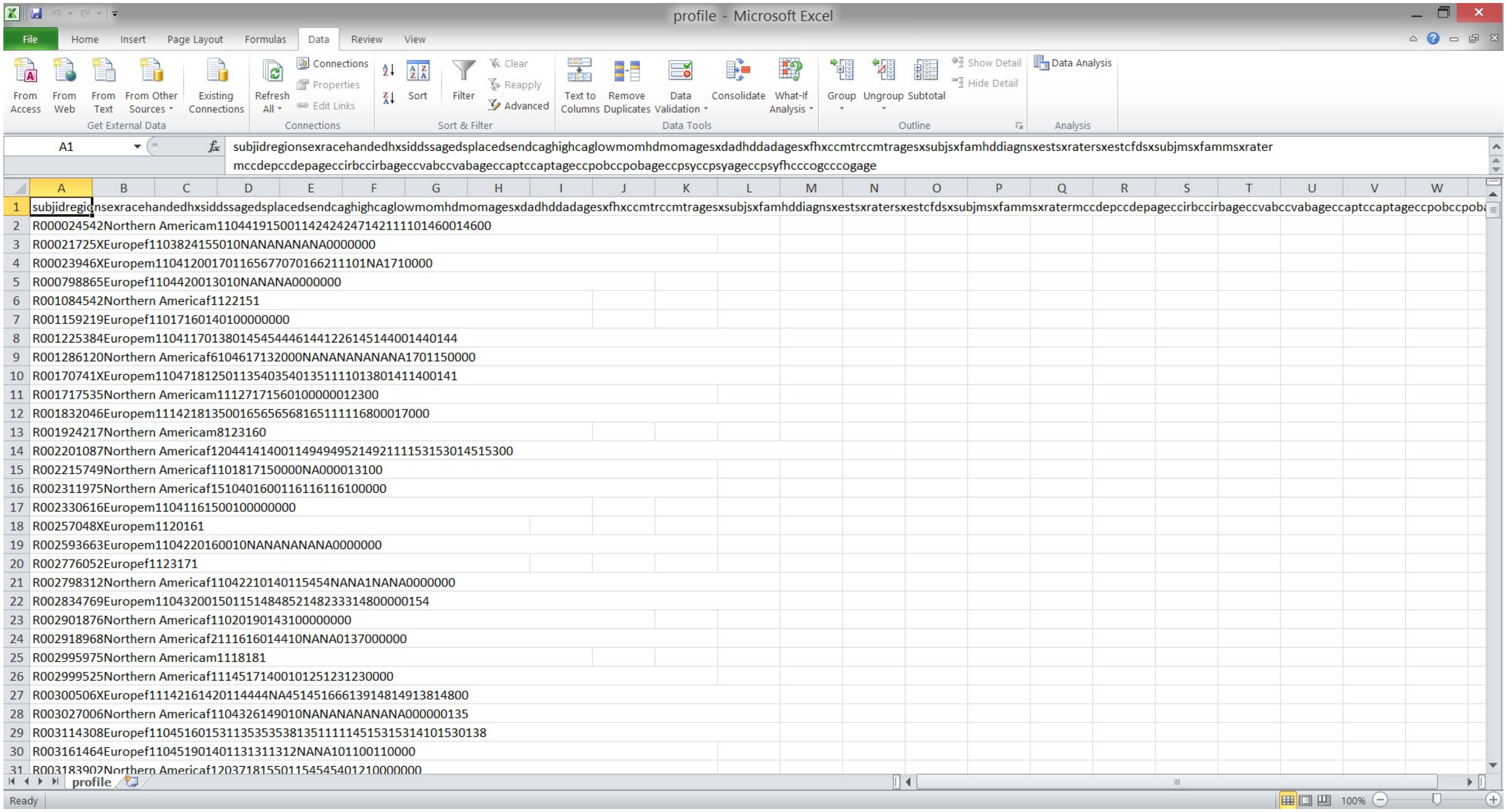
```
load(“Rdata_directory”)
```

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.

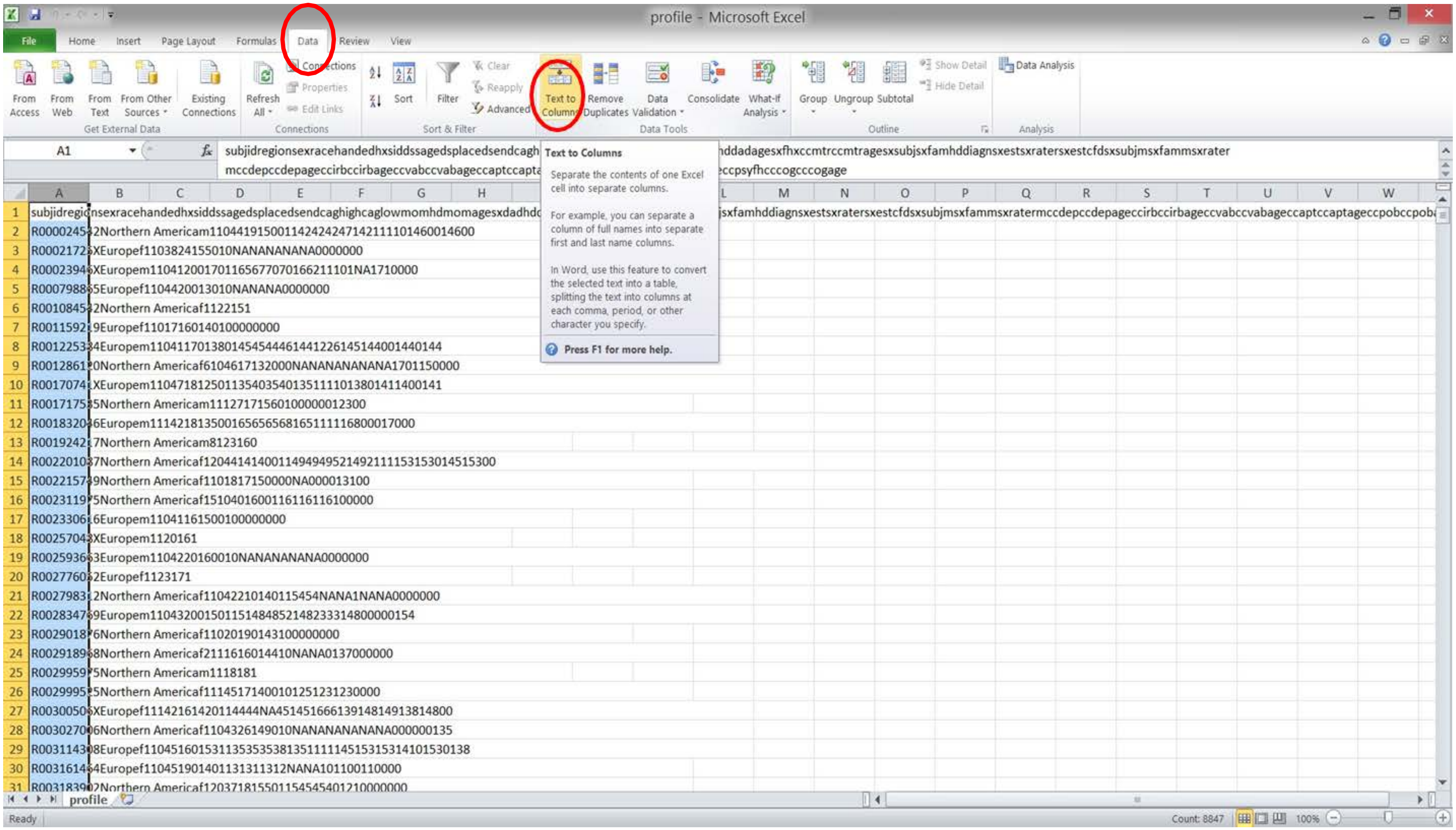
Step 1 – Open CSV file in Excel: Open the .csv file using Excel, or open Excel and on the “Data” tab click “**From Text/CSV**”. Data will be imported in entirety into the first column of the Excel file, as illustrated below.



```
A1  fx  subjidregionsexracehandedhxsiddssagedsplacedsendcaghighcaglowmomhdmomagesxdadhddadagesxfhxccmtrccmtragesxsubjsxfamhddiagnsxestsxratersxestcfdsxsubjmsxfammsxrater
        mccdepccdepageccirbccirbageccvabccvabageccaptccaptageccpobccpobageccpsyccpsyageccpsyfhcccogcccogage

     A
1    subjidregionsexracehandedhxsiddssagedsplacedsendcaghighcaglowmomhdmomagesxdadhddadagesxfhxccmtrccmtragesxsubjsxfamhddiagnsxestsxratersxestcfdsxsubjmsxfammsxratermccdepccdepageccirbccirbageccvabccvabageccaptccaptageccpobccpoba
2    R000024542Northern Americam11044191500114242424714211110146001460
3    R00021725XEuropef1103824155010NANANANANA0000000
4    R00023946XEuropem110412001701165677070166211101NA1710000
5    R000798865Europef1104420013010NANANA0000000
6    R001084542Northern Americaf1122151
7    R001159219Europef1101716014010000000 0
8    R001225384Europem110411701380145454446144122614514400144014 4
9    R001286120Northern Americaf6104617132000NANANANANANA1701150000
10   R00170741XEuropem11047181250113540354013511110138014114001 41
11   R001717535Northern Americam111271715601000000 12300
12   R001832046Europem111421813500165656568165111116800017000
13   R001924217Northern Americam8123160
14   R002201087Northern Americaf120441414001149494952149211115315301451530 0
15   R002215749Northern Americaf1101817150000NA000013100
16   R002311975Northern Americaf151040160011611611610000 0
17   R002330616Europem110411615001000000 00
18   R00257048XEuropem1120161
19   R002593663Europem110422016001 0NANANANANA0000000
20   R002776052Europef1123171
21   R002798312Northern Americaf11042210140115454NANA1NANA0000000
22   R002834769Europem1104320015011514848521482333148000001 54
23   R002901876Northern Americaf11020190143100000000
24   R002918968Northern Americaf2111616014410NANA0137000000
25   R002995975Northern Americam1118181
26   R002999525Northern Americaf11145171400101251231230000
27   R00300506XEuropef11142161420114444NA4514516661391481491381480 0
28   R003027006Northern Americaf1104326149010NANANANANANA000000135
29   R003114308Europef1104516015311353535381351111451531531410153013 8
30   R003161464Europef11045190140113131131 2NANA101100110000
31   R003183902Northern Americaf12037181550115454540121000000 0
```

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.



Text to Columns

Separate the contents of one Excel cell into separate columns.

For example, you can separate a column of full names into separate first and last name columns.

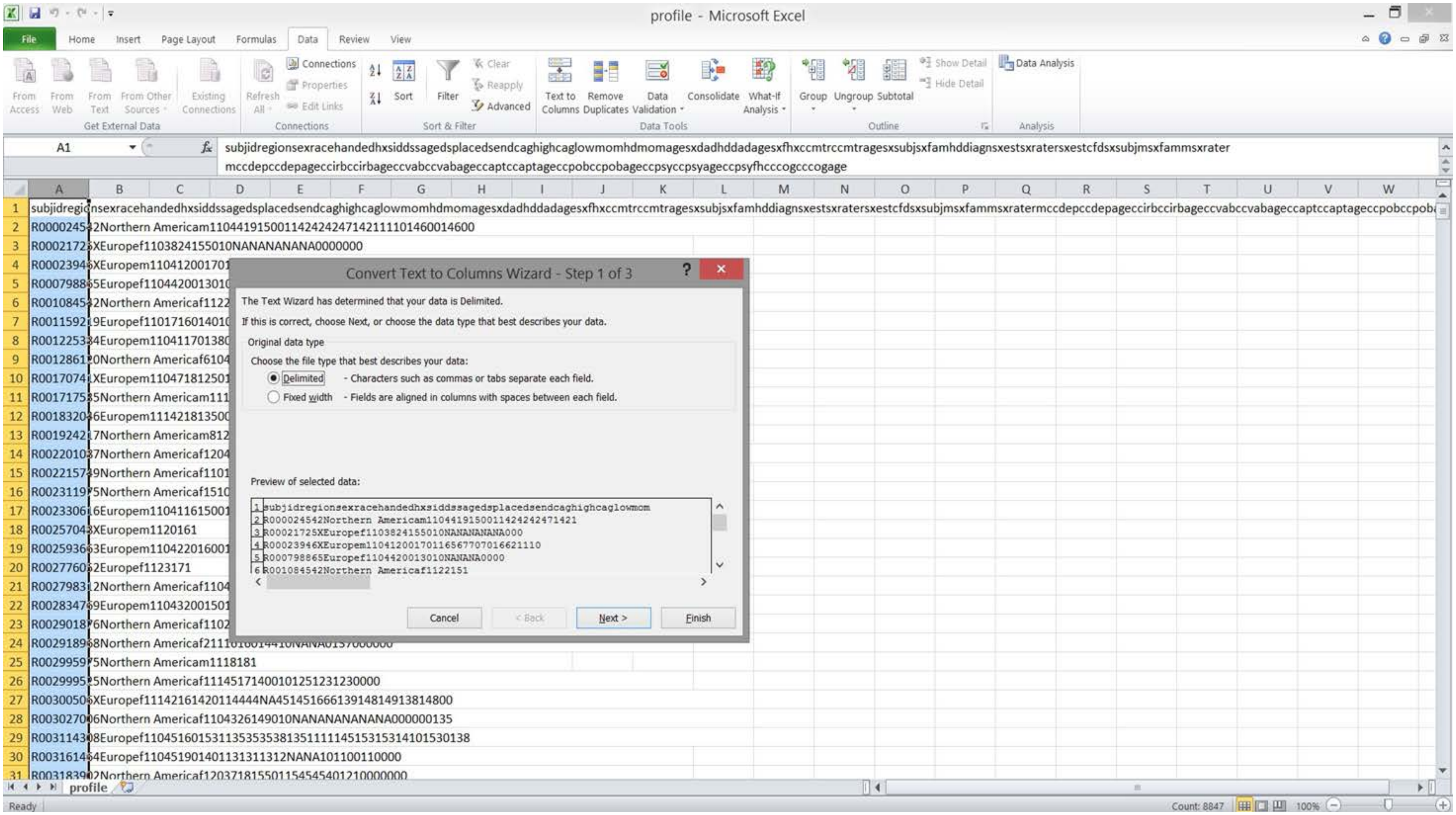
In Word, use this feature to convert the selected text into a table, splitting the text into columns at each comma, period, or other character you specify.

Press F1 for more help.

```
subjidregionsexracehandedhxsiddssagedsplacedsendcagh…hddadagesxfhxccmtrccmtragesxsubjsxfamhddiagnsxestsxratersxestcfdsxsubjmsxfammsxrater
mccdepccdepageccirbccirbageccvabccvabageccaptccapta…ccpsyfhcccogcccogage
```

```
1  subjidregionsexracehandedhxsiddssagedsplacedsendcaghighcaglowmomhdmomagesxdadhdc…jsxfamhddiagnsxestsxratersxestcfdsxsubjmsxfammsxratermccdepccdepageccirbccirbageccvabccvabageccaptccaptageccpobccpob
2  R00002452Northern Americam11044191500114242424714211110146001460
3  R00021725XEuropef1103824155010NANANANANA0000000
4  R00023946XEuropem11041200170116567707016621110 1NA1710000
5  R00079885 5Europef1104420013010NANANA0000000
6  R00108452Northern Americaf1122151
7  R00115921 9Europef1101716014010000000 0
8  R00122533 4Europem110411701380145454446144122614514400144014 4
9  R00128612 0Northern Americaf6104617132000NANANANANANA1701150000
10 R00170741XEuropem11047181250113540354013511110138014114001 41
11 R00171753 5Northern Americam11127171560100000012300
12 R00183204 6Europem111421813500165656568165111116800017000
13 R00192421 7Northern Americam8123160
14 R00220103 7Northern Americaf1204414140011494949521492111153153014515300
15 R00221574 9Northern Americaf1101817150000NA000013100
16 R00231197 5Northern Americaf151040160011611611610000 0
17 R00233061 6Europem110411615001000000 00
18 R00257043XEuropem1120161
19 R00259366 3Europem110422016001 0NANANANANA0000000
20 R00277606 2Europef1123171
21 R00279831 2Northern Americaf110422101401154 54NANA1NANA0000000
22 R00283476 9Europem1104320015011514848521482333148000001 54
23 R00290187 6Northern Americaf1102019014310000000 0
24 R00291896 8Northern Americaf211161601441 0NANA0137000000
25 R00299597 5Northern Americam1118181
26 R00299952 5Northern Americaf111451714001012512312300 00
27 R00300506XEuropef111421614201144 44NA45145166613914814913814800
28 R00302700 6Northern Americaf110432614901 0NANANANANANA000000135
29 R00311430 8Europef11045160153113535353813511111451531531410153013 8
30 R00316146 4Europef110451901401131311312NANA101100110000
31 R00318390 2Northern Americaf12037181550115454540121000000 0
```

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**".



profile - Microsoft Excel

A1: subjidregionsexracehandedhxsiddssagedsplacedsendcaghighcaglowmomhdmomagesxdadhddadagesxfhxccmtrccmtragesxsubjsxfamhddiagnsxestsxratersxestcfdsxsubjmsxfammsxrater mccdepccdepageccirbccirbageccvabccvabageccaptccaptageccpobccpobageccpsyccpsyageccpsyfhcccogcccogage

Convert Text to Columns Wizard - Step 1 of 3

The Text Wizard has determined that your data is Delimited.

If this is correct, choose Next, or choose the data type that best describes your data.

Original data type

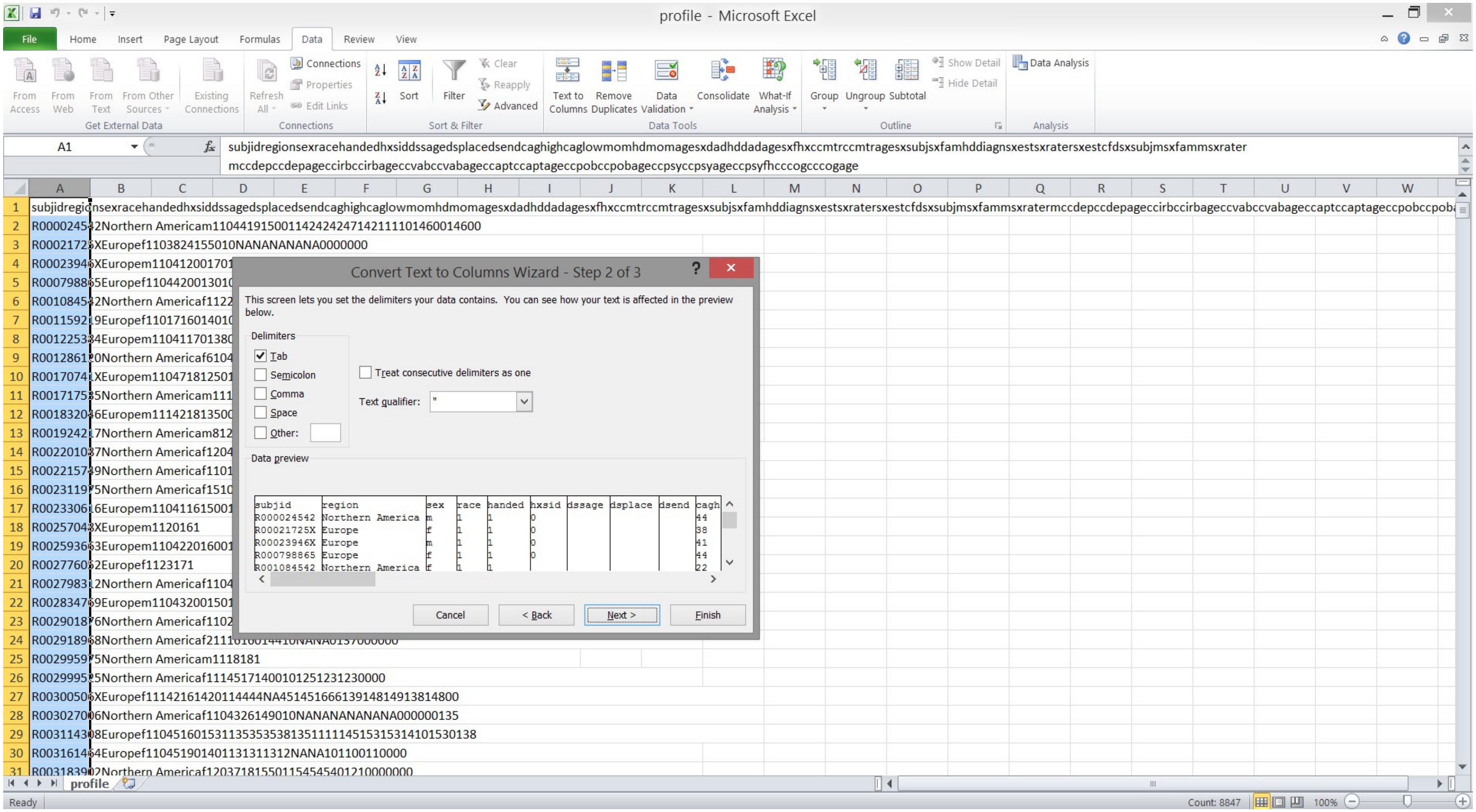
Choose the file type that best describes your data:

- ● Delimited - Characters such as commas or tabs separate each field.
- ○ Fixed width - Fields are aligned in columns with spaces between each field.

Preview of selected data:

```
1 subjidregionsexracehandedhxsiddssagedsplacedsendcaghighcaglowmom
2 R000024542Northern Americam11044191500114242424714 21
3 R00021725XEuropef1103824155010NANANANANA000
4 R00023946XEuropem11041200170116567707016621110
5 R000798865Europef1104420013010NANANA0000
6 R001084542Northern Americaf1122151
```

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**".



subjidregionsexracehandedhxsiddssagedsplacedsendcaghighcaglowmomhdmomagesxdadhddadagesxfhxccmtrccmtragesxsubjsxfamhddiagnsxestsxratersxestcfdsxsubjmsxfammsxrater mccdepccdepageccirbccirbageccvabccvabageccaptccaptageccpobccpobageccpsyccpsyageccpsyfhcccogcccogage

Convert Text to Columns Wizard - Step 2 of 3

This screen lets you set the delimiters your data contains. You can see how your text is affected in the preview below.

Delimiters

☑ Tab
☐ Semicolon
☐ Comma
☐ Space
☐ Other:

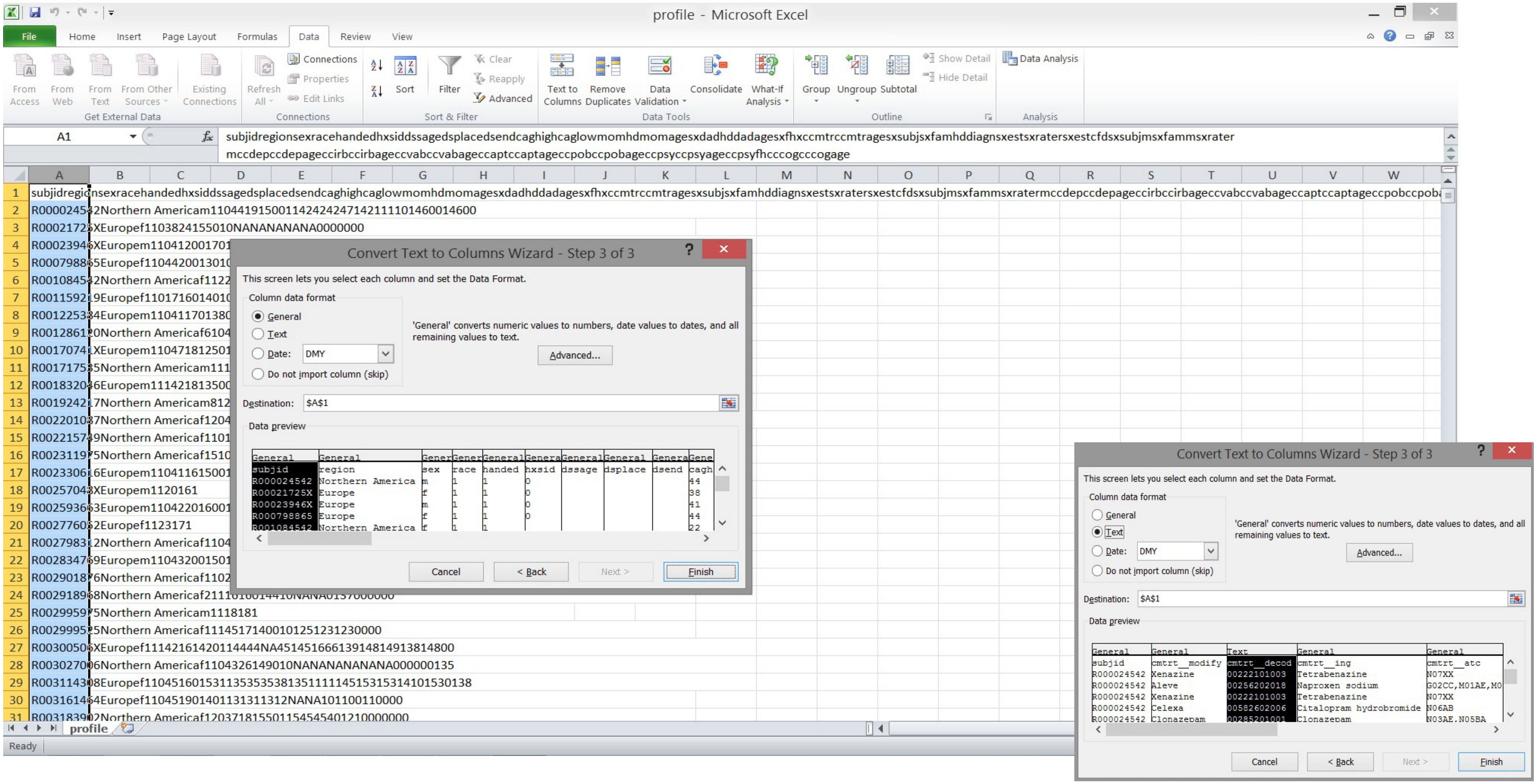
☐ Treat consecutive delimiters as one

Text qualifier: "

Data preview

```
subjid     region            sex  race handed hxsid dssage dsplace dsend cagh
R000024542 Northern America  m    1    1      0                          44
R00021725X Europe            f    1    1      0                          38
R00023946X Europe            m    1    1      0                          41
R000798865 Europe            f    1    1      0                          44
R001084542 Northern America  f    1    1                                 22
```

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**".



```
subjidregionsexracehandedhxsiddssagedsplacedsendcaghighcaglowmomhdmomagesxdadhddadagesxfhxccmtrccmtragesxsubjsxfamhddiagnsxestsxratersxestcfdsxsubjmsxfammsxrater
mccdepccdepageccirbccirbageccvabccvabageccaptccaptageccpobccpobageccpsyccpsyageccpsyfhcccogcccogage
```

```
1  subjidregionsexracehandedhxsiddssagedsplacedsendcaghighcaglowmomhdmomagesxdadhddadagesxfhxccmtrccmtragesxsubjsxfamhddiagnsxestsxratersxestcfdsxsubjmsxfammsxratermccdepccdepageccirbccirbageccvabccvabageccaptccaptageccpobccpobageccpsyccpsyageccpsyfhcccogcccogage
2  R000024542Northern Americam11044191500114242424714211110146001460 0
3  R00021725XEuropef1103824155010NANANANANA0000000
4  R00023946XEuropem110412001701
5  R000798865Europef110442001301O
6  R001084542Northern Americaf1122
7  R001159219Europef110171601401O
8  R001225334Europem110411701380
9  R001286120Northern Americaf6104
10 R00170741XEuropem110471812501
11 R001717535Northern Americam111
12 R001832046Europem111421813500
13 R001924217Northern Americam812
14 R002201037Northern Americaf1204
15 R002215749Northern Americaf1101
16 R002311975Northern Americaf1510
17 R002330616Europem110411615001
18 R00257043XEuropem1120161
19 R002593663Europem110422016001
20 R002776052Europef1123171
21 R002798312Northern Americaf1104
22 R002834769Europem110432001501
23 R002901876Northern Americaf1102
24 R002918968Northern Americaf2111616014410NANA0137000000
25 R002995975Northern Americam1118181
26 R002999525Northern Americaf1114517140010125123123 0000
27 R00300506XEuropef11142161420114444NA4514516661391481491381480 0
28 R003027006Northern Americaf1104326149010NANANANANANA000000135
29 R003114308Europef1104516015311353535381351111145153153141015301 38
30 R003161464Europef1104519014011313113 12NANA101100110000
31 R003183902Northern Americaf12037181550115454540121000 0000
```

Convert Text to Columns Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

- ● General
- ○ Text
- ○ Date: DMY
- ○ Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Advanced...

Destination: A1

Data preview

```
General     General           GenerGenerGeneralGeneraGeneralGeneral GeneraGene
subjid      region            sex  race handed hxsid dssage dsplace dsend cagh
R000024542  Northern America  m    1    1      0                          44
R00021725X  Europe            f    1    1      0                          38
R00023946X  Europe            m    1    1      0                          41
R000798865  Europe            f    1    1      0                          44
R001084542  Northern America  f    1    1                                 22
```

Cancel | < Back | Next > | Finish

Convert Text to Columns Wizard - Step 3 of 3

This screen lets you select each column and set the Data Format.

Column data format

- ○ General
- ● Text
- ○ Date: DMY
- ○ Do not import column (skip)

'General' converts numeric values to numbers, date values to dates, and all remaining values to text.

Advanced...

Destination: A1

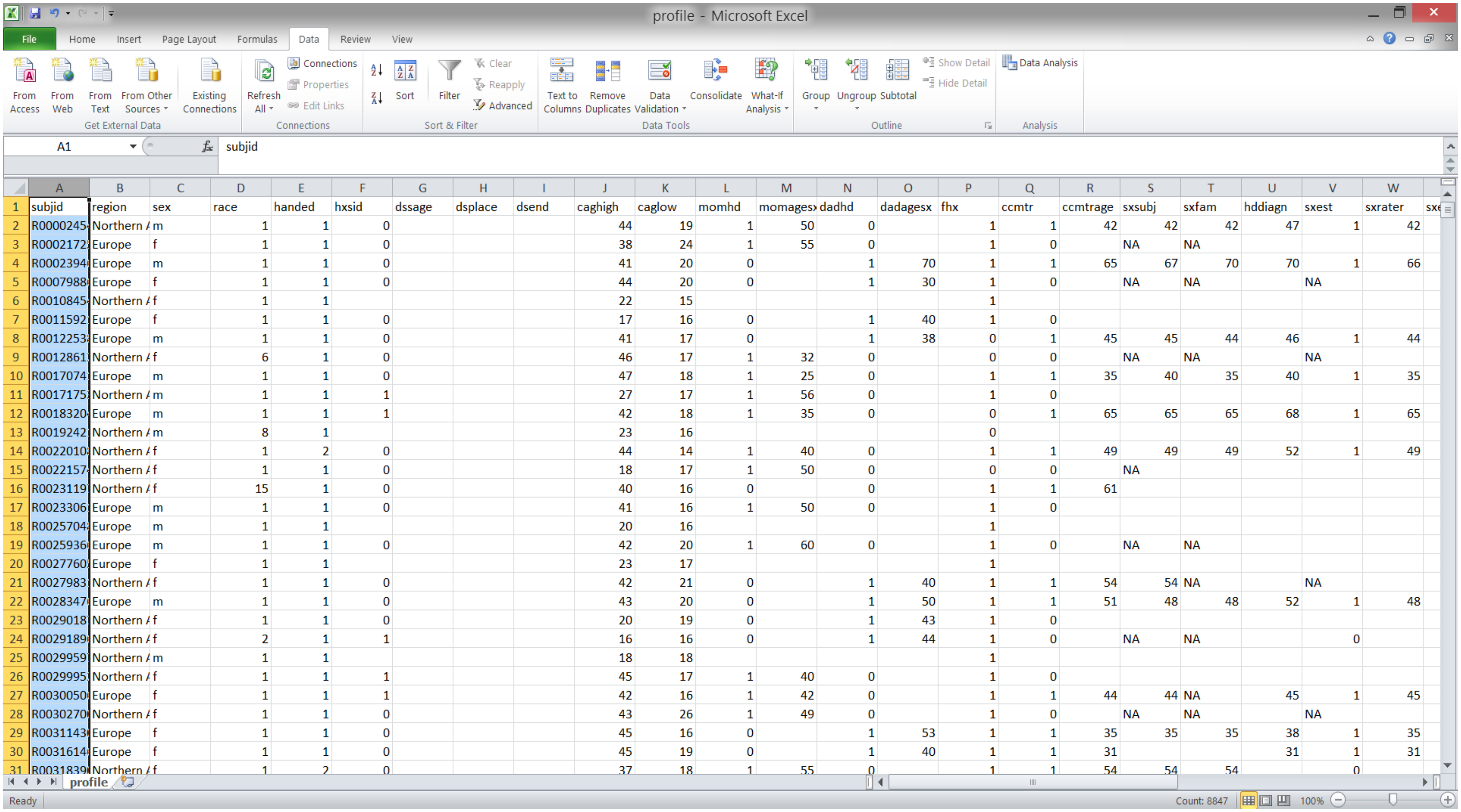
Data preview

```
General     General         Text          General                   General
subjid      cmtrt__modify   cmtrt__decod  cmtrt__ing                cmtrt__atc
R000024542  Xenazine        00222101003   Tetrabenazine             N07XX
R000024542  Aleve           00256202018   Naproxen sodium           G02CC,M01AE,M0
R000024542  Xenazine        00222101003   Tetrabenazine             N07XX
R000024542  Celexa          00582602006   Citalopram hydrobromide   N06AB
R000024542  Clonazepam      00285201001   Clonazepam                N03AE,N05BA
```

Cancel | < Back | Next > | 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.



Cell A1: subjid

Count: 8847

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