SEER data pull method

1. Download Seer\*Stat software ( <https://seer.cancer.gov/seerstat/software/>). Note: must register to download the software.
2. Create an account to access the data ( <https://seer.cancer.gov/data/access.html> )

* It will take ~1 hour to be approved for an account.
* You will receive separate emails with username and password.

1. Open Seer\*Stat (only available on Windows, so if you have a Mac use Parallels (<https://www.parallels.com>) or another Windows emulator).
2. Choose File -> New -> Case Listing Session

Graphical user interface, application, Word

Description automatically generated

1. Highlight “Incidence – SEER Research Limited-Field Data, 22 Registries, Nov 2021 Sub (2000-2019)” in Data tab

Graphical user interface, text, application

Description automatically generated

1. In the Table tab, highlight “Year of Diagnosis” from the ‘Race, Sex, Year Dx’ folder and select Column to add as column. Do the same for “Diagnostic Confirmation”, “Histologic Type ICD-0-3”, “ICD-0-3 Hist/behav” (all within the ‘Site and Morphology’ folder), and “Patient ID” (in the ‘Other’ folder) as columns

Graphical user interface

Description automatically generated

1. In the Selection tab, select ‘Edit…’on the righthand side of the window. In the new window, under ‘Variable’ open folder “Race, Sex, Year Dx” and highlight “Year of diagnosis”. For ‘Operator’ highlight “is = to”, and for ‘Values’ click and drag to highlight ‘2010 – 2019’ only. Select OK.

Graphical user interface, application

Description automatically generated

1. Press the execute button (lightning bolt).

Graphical user interface, text, application

Description automatically generated

A popup should appear showing the progress of the download:

A picture containing graphical user interface

Description automatically generated

1. Once the data finishes processing you should get another popup of the matrix, to export this as a text file choose Matrix (in the top banner) -> Export -> ‘Results as Text File…’ and an Export popup appears.
2. Name the file what you would like in ‘Data File’, and the dictionary file similarly in ‘Export Dictionary File’. Change ‘Missing Character’ to “NA”, ‘Field Delimiter’ to “Comma” and check the “Output Variable Names Before Data” checkbox. It should look as it does below. Press OK. Export may take 5+ minutes.

Graphical user interface, text, application, email

Description automatically generated

TO CHECK DATA PULL IS CONCORDANT

1. Download R and RStudio, (<https://courses.edx.org/courses/UTAustinX/UT.7.01x/3T2014/56c5437b88fa43cf828bff5371c6a924/> lays out the steps well)
2. Open RStudio, File -> New File -> R Markdown. You may get a pop-up window ‘Install Required Packages.’ Select Yes. This may take ~2 minutes
3. Name the file, press OK. In the Rmd. File that opens in the Source pane, delete all text that is highlighted below

Graphical user interface, text, application, email

Description automatically generated

1. Insert a new code chunk by clicking the green C at top of field.

Graphical user interface, application, Word

Description automatically generated

1. It should look like:

Graphical user interface, application

Description automatically generated

1. To import the text file, File -> Import Dataset -> From Text (base)… and choose the appropriate file that we saved above (from above SEER\_21registries\_data.txt )

Graphical user interface

Description automatically generated

1. A popup will appear with the heading “Import Dataset” that shows a subset of the data that was selected above. Next to Heading, choose Yes. Then click Import at the bottom of the popup.

Graphical user interface, application

Description automatically generated

1. The data set will appear in the Environment and the code needed will appear in the Console (shown below). Copy that code into the code chunk we inserted into the .Rmd file in step 14. (This is a way to bypass the issue of knowing the exact location of the .txt file on your machine.)

Graphical user interface, text, application

Description automatically generated

Line of code to copy into the .Rmd file above

Copy code from below into this code chunk

Dataset to view in the Environment

1. There are some packages that need to be installed onto the machine to make the following code chunk work, copy these lines into the Console one line at a time:

install.packages("magrittr")

install.packages("dplyr")

install.packages("janitor")

install.packages(“readr”)

1. After these packages have been installed, they need to be referenced within the .Rmd file. Copy and paste these lines into the top code chunk (starting “knitr::opts\_chunk$set(echo = TRUE)”).

library(magrittr)

library(dplyr)

library(janitor)

library(readr)

Graphical user interface, text, application

Description automatically generated

1. In the code chunk where we imported the dataset, copy and paste these lines of code below the read.csv() line from step 18.

#import data and clean names

SEER\_22registries\_data <- SEER\_22registries\_data %>%

rename(ICD = ICD.O.3.Hist.behav) %>%

clean\_names()

# separate solid and heme at 9590/3 where heme includes 9590/3

solid\_yr22 <- SEER\_22registries\_data %>%

filter(histologic\_type\_icd\_o\_3 < "9590/3")

heme\_yr22 <- SEER\_22registries\_data %>%

filter(histologic\_type\_icd\_o\_3 >= "9590/3")

1. Create a new code chunk (step 14) and copy and paste this code in it to check if the data is the same (through a simple check of number of cases per year for both solid and heme)

solid\_yr22 %>%

group\_by(year\_of\_diagnosis) %>%

summarise(n = n())

heme\_yr22 %>%

group\_by(year\_of\_diagnosis) %>%

summarise(n = n())

The output tables should look like:

SOLID HEME

Table

Description automatically generated Table

Description automatically generated