Data Analysis on SEER cancer data

University of Wisconsin-Milwaukee

Venkat Hemant Akurati

Abstract

Background: I Acquired a Dataset from The National Cancer Institute's SEER (Surveillance, Epidemiology and End Results) Program conducted in 28 percent of Us population on different cancer incidences which is funded by NCI and CDC. Using the Data set from SEER Program I tried to find out the cancer occurrences in all types in different age groups ranges for both men and women Respectively.

Method: Firstly I Obtained the SEER dataset by signing and sending the user-agreement form at http://seer.cancer.gov/data/request.html . Then I obtained the login-password to access the SEER dataset which I downloaded from http://seer.cancer.gov/data/options.html the ASCII text version of the data and uncompressed it. Later I saved a file with ICD-O3 codes with the respective diseases. I performed data analytics on the data and saved an output results file in .csv format for future references.

Results: I saved a datafile with output data after analysis which consists of number of cases with different cancers for different age groups and genders. Based on the results obtained majority of the cases reported are among the men(age>75) and adenocarcinoma ranks among the highest.

Introduction

National Institute of Health (NIH) is a part of U.S Department of health and human services. NIH is the steward of medical and behavioral research for the Nation. Its mission is to seek fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability. [1] Surveillance, Epidemiology, and End Results (SEER) Program provides information on cancer statistics in an effort to reduce the cancer burden among the U.S. population. SEER is supported by the Surveillance Research Program (SRP) in NCI's Division of Cancer Control and Population Sciences (DCCPS). [2] Information on cancer incidences is collected from areas representing 28 percent of the US population, funded by NCI and CDC. SEER Collects data from Approximately 34.6 percent population from the Unites States. The cancer Incidences are recorded from Population-based cancer registries. SEER research data is a collection of over 10 million cancer de-identified records of incidences that occurred between 1974-2016 (2019 release). This dataset is updated every year. To access the data a data request form should be submitted stating the purpose. The data is based on Patient demographics, primary tumor site, tumor morphology, stage at diagnosis, and first course of treatment. All the Cases reported in SEER database are based on the ICD -O-3 histology and behavior code. ICD -O-3 is designed to categorize tumors reference from International Classification of Diseases for Oncology, third edition. Used principally in tumor or cancer registries for coding the site (topography) and the histology (morphology) of neoplasms, usually obtained from a pathology report. [3]

Creating an ICD-O Dictionary

There is a file in text format that contains ICD codes and their relative disease names available freely in the SEER website at the following link (http://seer.cancer.gov/icd-o-3/). I created an ICD-

O Dictionary assigning key and value pairs where keys are the codes of the diseases and values as the name of the disease. Below is the code snippet.

```
import re

# A function that makes a dictionary from ICD-O codes;
# codes as keys and disease name as values
def ICDO_dictionary():

    d = {}

    infile = open("icdo3.txt","r")
    for line in infile:
        # regular expression to read the codes and dieseas names
        r = re.search("(\d+/\d+)\s+(.+)",line)

        if (r) :
            code = r.group(1)
            code = code.replace("/","")
            disease = r.group(2)
            disease=disease.rstrip()
            disease = disease.lower()
            d[code] = disease

    infile.close()
    return d
```

Data Description

After receiving the access from the NCI dataset can be downloaded. The data in the directory are stored in ASCII text format. This TEXTDATA directory contains the following

- SEER Cancer Incidence Research Database for patients diagnosed 1975-2016
- County level population estimates for the SEER registries by 19 age groups and single ages from the Census Bureau for 1975-2016.
- Data dictionaries for the SEER and population data \INCIDENCETEXTDATA.FILEDESCRIPTION.PDF - Data dictionary for the SEER incidence data files.

Incidence folder has data from several registries stored under 4 folders. Each folder has 9 files categorized based on the anatomical sites of cancer. Each line is a record of cancer case described by alphanumeric characters.

BREAST.TXT - Breast

COLRECT.TXT - Colon and Rectum

DIGOTHR.TXT - Other Digestive

FEMGEN.TXT - Female Genital

LYMYLEUK.TXT - Lymphoma of All Sites and Leukemia

MALEGEN.TXT - Male Genital

RESPIR.TXT - Respiratory

URINARY.TXT - Urinary

OTHER.TXT - All Other Sites

Data Analysis

I want to count the number of cases with cancers for different age group ranges between males and females. This code takes the names of the diseases from the ICD-O dictionary based on the key values from the SEER dataset. The result output data is saved in the comma separated value(CSV) format for future manipulations, sorting and graphs. Below are the code snippets.

```
import csv
import re


def main():
    d = {}
    Infile = open("icdo3.txt", "r")
    for line in Infile:
        r = re.search("(\d+/\d+)\s+(.+)", line)
        if r:
            code = r.group(1)
            code = code.replace("/", "")
            disease = r.group(2)
            disease = disease.rstrip()
            disease = disease.lower()
            d[code] = disease
    Infile.close()

    filelist = glob.glob("C:/Users/avula/Desktop/Priyanka/Assignment-10/SEER_1975_2016_TEXTDATA/incidence/yr*/*.TXT")
    # print(filelist)
    # filelist=glob.glob("C:/Users/avula/Desktop/Priyanka/Assignment-10/test.txt")
    di = {}  # dictionary with ICD-O codes as keys and # of occurrences as values
    for file in filelist:  # process each file
        # print(file)
        infile = open(file, "r")
        for line in infile:
            disease = line[52:57]  # ICD-Oncology-3 code (Histology+Behavior) )
            gender = line[23]
            birthstring = line[27:31]
            if (birthstring == "    "):
                birth = 0
            else:
                birth = int(birthstring)
            age = 2019 - birth
            # print(disease)
            # print(gender)
            # print(age)
            # index is the location in the list which needs to be updated
            if gender == "1":  # men
                if age < 25:
                    index = 0
                    # print("men25")
                elif age >= 25 and age < 50:
                    index = 2
                    # print("men25-50")
                elif age >= 50 and age < 75:
                    index = 4
                    # print("men50-75")
                elif age > 75:
                    index = 6
                    # print("men75+")
            if gender == "2":  # women
                if age < 25:
                    index = 1
```

```
                    # print("women25")
                elif age >= 25 and age < 50:
                    index = 3
                    # print("women25-50")
                elif age >= 50 and age < 75:
                    index = 5
                    # print("women50-75")
                elif age > 75:
                    index = 7
                    # print("women75+")
            if disease in di:
                di[disease][index] += 1
            else:
                di[disease] = [0] * 8  # list of eight zeros
                di[disease][index] += 1
            # a = int(line[52:57]) # ICD-Oncology-3 code (Histology+Behavior) appears at 53-57 characters (then why 52-57?)
            # di[a] = di.get(a,0)+1 # increment the count in the dictionary
        infile.close()
    # print(di)
    #print(d)

    w = csv.writer(open("output1.csv", "w"))
    for key, val in di.items():
        # print(val[0],val[1],val[2],val[3],val[4],val[5],val[6],val[7])
        name = str((d.get(key)))
        name = name.replace(",", "")
        w.writerow([name, val[0], val[1], val[2], val[3], val[4], val[5], val[6], val[7]])


main()
```

Results

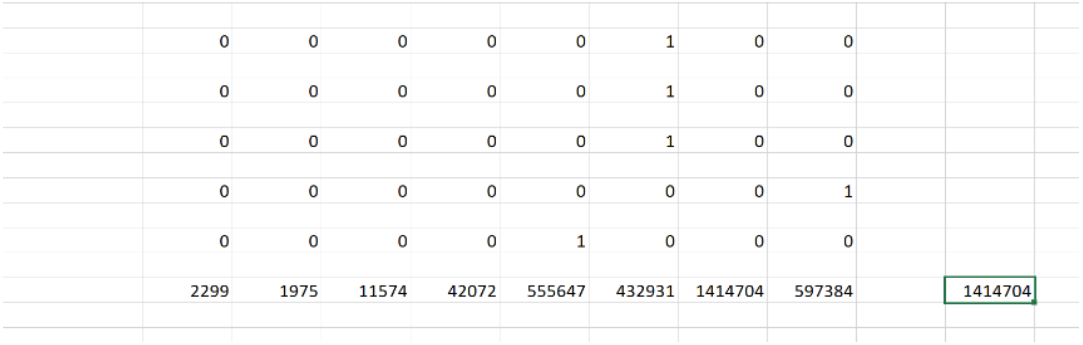
The result output data file shows the result as disease name and the number of cases based on the age and gender. Below is the sample snapshot of the code.

```
duct carcinoma in situ solid type,0,0,0,701,25,9002,11,4820
intraductal carcinoma noninfiltrating nos,2,0,16,2663,179,51022,326,51010
infiltr. duct mixed with other types of carcinoma in situ,0,0,9,3171,102,36558,87,19279
carcinoma nos,28,23,977,2195,22495,24577,144688,142616
intracystic carcinoma nos,0,0,6,26,33,371,52,1072
lobular carcinoma nos,0,0,0,1978,41,45655,81,67624
adenocarcinoma nos,98,41,10725,11096,555647,178999,1414704,597384
infiltrating duct and lobular carcinoma,0,0,3,2190,57,37761,97,42150
paget disease and intraductal ca.,0,0,0,57,3,794,12,1950
```

Explanation of the result data the first part consists of the disease name followed by as follows Men (age <25), Women (age <25), Men (age 25-50), Women (age 25-50), Men (age a50-75), Women (age 50-75), Men (age >75), Women (age >75).

Based on the results obtained

1. Highest average mean value of the all the cases recorded is among the age group of Men(age>75).
2. Total number of the cancer cases reported are also higher among the age group of >75 and particularly among men which is 3412528 cases.
3. Highest number of cases recorded is adenocarcinoma nos among men(age>75) = 1414704.



0	0	0	0	0	1	0	0	
0	0	0	0	0	1	0	0	
0	0	0	0	0	1	0	0	
0	0	0	0	0	0	0	1	
0	0	0	0	1	0	0	0	
2299	1975	11574	42072	555647	432931	1414704	597384	1414704

References

1. About the NIH. (2015). Retrieved from https://www.nih.gov/about-nih/what-we-do/nih-almanac/about-nih
2. Surveillance, Epidemiology, and End Results Program. (2020). Retrieved from https://seer.cancer.gov/
3. International Classification of Diseases for Oncology, 3rd Edition (ICD-O-3). (2020). Retrieved from https://www.who.int/classifications/icd/adaptations/oncology/en/
4. Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Research Data (1975-2016), National Cancer Institute, DCCPS, Surveillance Research Program, released April 2019, based on the November 2018 submission.