Analysis Statistics of United States Cancer

Xiaoxu Na

Abstract –The dread and fear that can come with a cancer diagnosis have their roots in its killer nature: It's the No. 2 cause of death in Americans, second only to heart disease, according to the Centers for Disease Control and Prevention. Even when diagnosed early and attacked with the latest treatments, it still has the power to kill. This project forces statistical analysis of the cancers in the U.S.

Index Term – United States Cancer, analysis, RStudio, machine learning

1 Introduction

About one-third of all people in the US will develop cancer during their lifetimes. Over one and a half million new cancer cases are diagnosed each year. Anyone can get cancer at any age, but the risk goes up with age. Nearly 9 out of 10 cancers are diagnosed in people ages 50 and older. Cancer can be found in people of all racial and ethnic groups, but the rate of cancer occurrence (called the incidence rate) varies from group to group. Today, more than 15 mil-lion people alive in the United States have had some type of cancer. Some of these people are cancer-free; others still have it. [1] This project will look closely at what happens in large groups of people and provide a picture in time of the bur-den of cancer on society. The statistics also tell us about differences among groups defined by age, sex, racial/ethnic group, and other categories.

2 Backgrounds

2.1 A Collection of Related Diseases

Cancer is the name given to a collection of related diseases. In all types of cancer, some of the body's cells begin to divide without stopping and spread into surrounding tissues. When cancer develops, as cells become more and more abnormal, old or damaged cells survive when they should die, and new cells form when they are not needed. These extra cells can divide without stopping and may form growths called tumors. Cancerous tumors are malignant, which means they can spread into, or invade, nearby tissues. In addition, as these tumors grow, some cancer cells can break off and travel to distant places in the body through the blood or the lymph system and form new tumors far from the original tumor. [2]

2.2 Types of Cancer

There are more than 100 types of cancer. Types of cancer are usually named for the organs or tissues where the cancers form. For example, lung cancer starts in cells of the lung, and brain cancer starts in cells of the brain. They may also be described by the type of cell that formed them, such as an epithelial cell or a squamous cell. This project will take the types as indicated in the dataset, such as breast cancer, brain and other nervous can-cer, stomach cancer and etc.

2.3 RStudio

RStudio is an integrated development environment (IDE) for R. It includes a con-sole, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management. RStudio is available in open source and commercial editions and runs on the desktop or in a browser connected to RStudio Server or RStudio Server. [3] It is very convenient to do machine learning in R as well.

3 Dataset

The data used in the project comes from Data Citation1: United States Cancer Statistics (USCS) published on the website of Center for Disease Control and Prevention. [1]

4 Dataset Analysis

4.1 The Cancer Mortality Rate of Female

4.1.1 The Overall View of Cancer Mortality Rate of Female

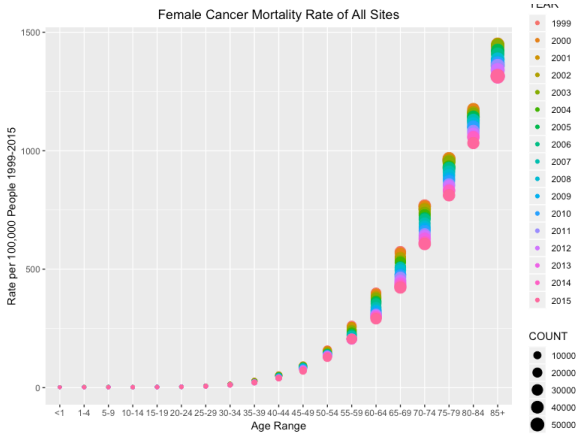

Figure 1.1

From the fig 1.1, we may tell that the female cancer mortality rate of all sites tends to go down from the year 1999 to 2015 for the groups of all different age ranges.

4.1.2 The Cancer Mortality Rate of Female based on Race

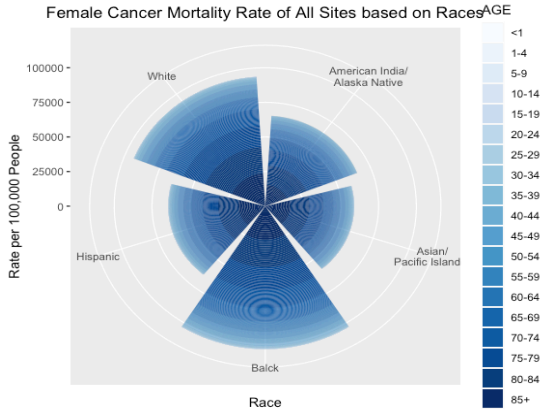

Figure 1.2

Based on fig 1.2, it is indicated that the cancer mortality rate of female varies on different races. The statistics information is divided into five groups. The morality rates of three groups are almost the same, however, the mortality rates of black and white people are higher than the rest people.

4.1.3 The Cancer Mortality Rate of Female based on Cancer Sites

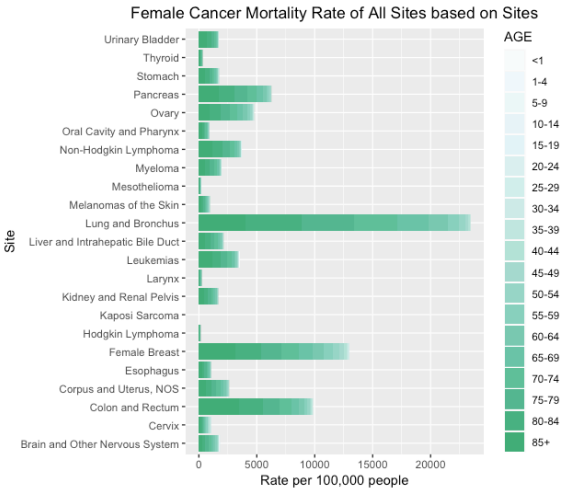

Figure 1.3

Based on fig 1.3, the lung and bronchus cancer caused the most death, and follow the female breast and colon and rectum cancers. These three types have a much higher mortality rate on females. The least female cancer mortality rates are with the type of kaposi sarcoma, hodgkin lymphoma and mesothelioma.

4.2 The Cancer Mortality Rate of Male

4.2.1 The Overall View of Cancer Mortality Rate of Male

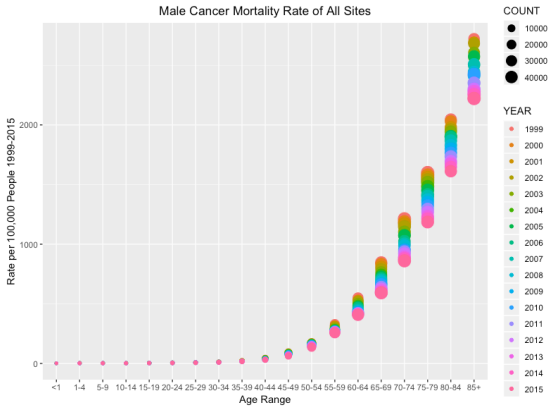

Figure 2.1

From the fig 2.1, we may tell that the male cancer mortality rate of all sites tends to go down from the year 1999 to 2015 for the groups of all different age ranges, especially for older males.

4.2.2 The Cancer Mortality Rate of Male based on Race

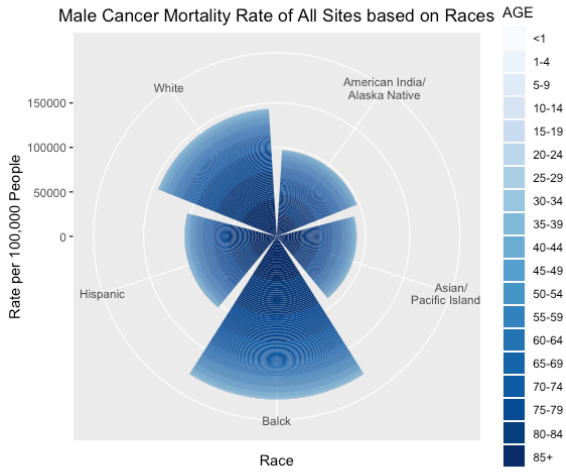

Figure 2.2

Based on fig 2.2, it is indicated that the cancer mortality rate of male varies on different races. The statistics information is also divided into five groups. The mortality rate of black people is higher than the rest.

4.2.3 The Cancer Mortality Rate of Male based on Cancer Sites

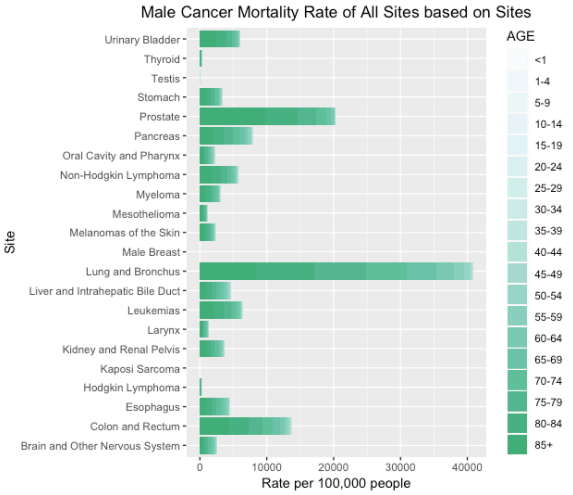

Figure 2.3

Concerning fig 2.3, the lung and bronchus cancer caused the most death, which is the same as that of females. Then follow the prostate and colon and rectum cancers. These three types have a much higher mortality rate on males. For males and females, as displayed on fig 1.3 and fig 2.3, the cancer mortality rates of the lung and bronchus and colon and rectum are the highest. The least male cancer mortality rates are with the type of kaposi sarcoma, male breast and testis.

4.3 The Cancer Incidence Rate of Female

4.3.1 The Overall View of Cancer Incidence Rate of Female

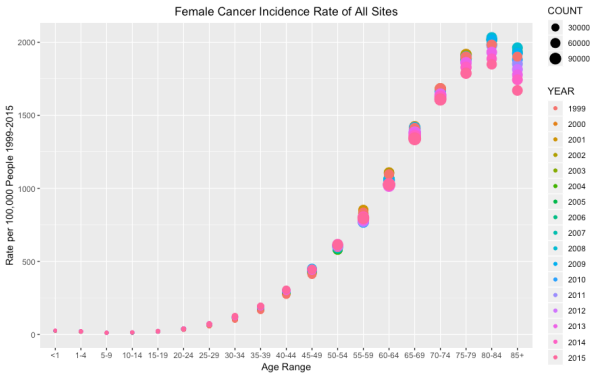

Figure 3.1

From the fig 3.1, we may tell that the female cancer incidence rate of all sites tends to go down a bit from the year 1999 to 2015 for the groups of all different age ranges.

4.3.2 The Cancer Incidence Rate of Female based on Race

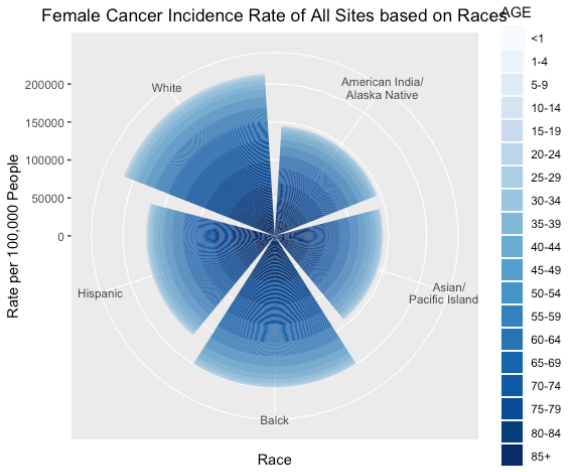

Figure 3.2

Based on fig 3.2, it is indicated that the cancer mortality rate of female varies a little on different races. The statistics information is divided into five groups. The morality rates of the five groups are almost the same.

4.3.3 The Cancer Incidence Rate of Female based on Cancer Sites

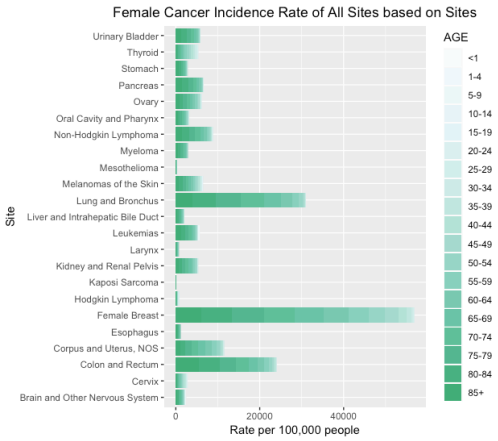

Figure 3.3

Concerning the fig 3.3, the female breast cancer happens quite a lot on females, and follows the lung and bronchus cancer and the colon and rectum cancer.

4.4 The Cancer Incidence Rate of Male

4.4.1 The Overall View of Cancer Incidence Rate of Male

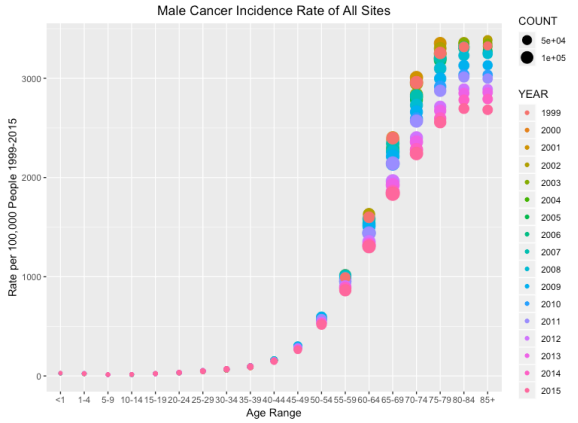

Figure 4.1

From the fig 4.1, we may tell that the male cancer incidence rates of all sites go down a lot from the year 1999 to 2015 for the groups of all different age ranges, especially for elder males.

4.4.2 The Cancer Incidence Rate of Male based on Race

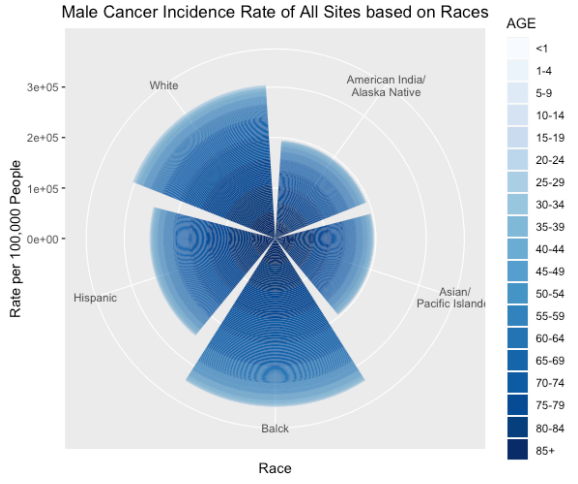

Figure 4.2

Based on fig 4.2, it is indicated that the cancer incidence rate of female varies a little on different races. The statistics information is divided into five groups. The morality rates of the five groups are nearly the same.

4.4.3 The Cancer Incidence Rate of Male based on Cancer Sites

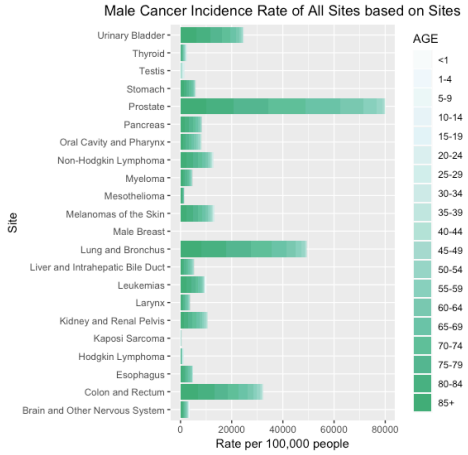

Figure 4.3

Concerning the fig 4.3, the prostate cancer happens quite a lot on males, and follows the lung and bronchus cancer and the colon and rectum cancer.

5 Prediction Results

5.1 The Female Cancer Mortality Rate of Age 85+ Prediction Model

Using linear regression method, it is constructed a linear model as shown in fig 5.1. This model took the group of 85+, and it is the same concept to predict other rate or count as well.

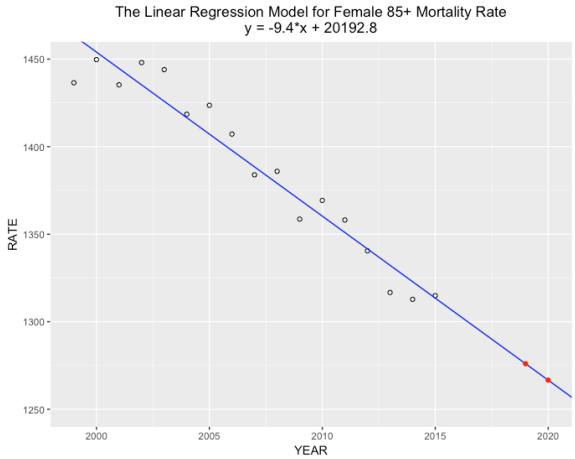

Figure 5.1

5.2 Display the Predicted Points on Previous Analysis

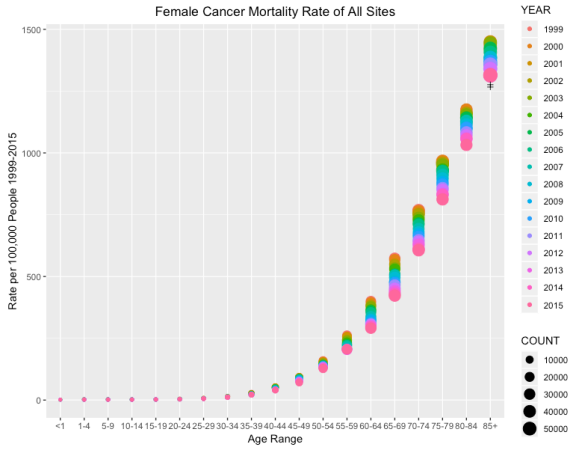

Figure 5.2

Since the new data were achieved from the prediction model, they were added to the previous graph as black points to show the tendency. The same concept works for other groups as well.

6 Future Works

There are many packages for prediction and R itself provides different prediction methods as well. It may be more accurate if the predictions are executed through different methods and having a compare among the methods, the best one may give the most accurate prediction.

Acknowledgements

The author would like to express the gratitude to Dr. Mary Yang.

Data Citations

[1] Download Data Tables: 1999-2015.zip, 9/13/2018.

https://www.cdc.gov/cancer/uscs/dataviz/download_data.htm

References

[1] “Brain Tumor: Statistics”, Approved by the Cancer Net Editorial Board, https://www.cancer.net/cancer-types/brain-tumor/statistics, 11/2017.
[2] “What is cancer?”, https://www.cancer.gov/about-cancer/understanding/what-is-cancer.
[3] “Take Control of Your R Code”, https://www.rstudio.com/products/RStudio/.