Examination of United States Cancer Mortality from 1999-2017 and trends between Race, Ethnicity, Age Group, etc.

Examining Cancer Mortality Through a Variety of Categorical Variables

Brent Cameron

2021-10-15

# 1 Summary/Abstract

I will analyze data to further understand trend of cancer mortality trends from 1999-2017 in the United States using data collected via the CDC wonder website. The goal of this project is to understand the role that age group, sex, ethnicity, and race play in the overall cancer trends from 1999-2017. This project will hopefully shed further light into what could be the best targeted approaches for minimizing disease burden for cancer in the future.

# 2 Introduction

## 2.1 General Background Information

The word cancer has long been one of the most recognizable and fear-inducing words worldwide in the sphere of chronic disease. Cancer, along with heart disease and diabetes, are responsible for the leading causes of death and disability in the United States (according to the Center for Disease Control-<https://www.cdc.gov/chronicdisease/about/index.htm#>:~:text=Chronic%20diseases%20such%20as%20heart,disability%20in%20the%20United%20States). Because of the vast scope and many different types of the disease, resources and networks have been continually strained in identifying the most efficient and effective ways to counteract cancerous cells. While a multitude of research has already been conducted on cancer, understanding the differences in how the disease continues to affect and interact with different populations can serve as an invaluable resource in tailoring new innovative strategies to combat the disease.

## 2.2 Description of data and data source

The data used in this project is a compilation of information regarding cancer mortality cases from 1999-2017 based on age group, sex, ethnicity. race, and cancer sites (\*note that this data only includes information collected from the United States). The data is composed of 8809 observations and comes from the CDC wonder website which can be found here:

\*<https://wonder.cdc.gov/cancermort-v2017.HTML>

## 2.3 Questions/Hypotheses to be addressed

Cancer has been and will continue to be widely studied- however, with the presence of constantly evolving data and new research shedding light on innovative techniques and approaches to better manage and control the disease, new examinations of data trends can provide a valuable framework for future studies. With that being said, there are several questions in particular that I will ask and attempt to answer through this research project:

\*Are cancer mortality trends dependent on or affected by race in the United States? If so, in what ways?

\*How does age factor into mortality trends of cancer? Will the past research concluding that the elderly are the most vulnerable continue to be held evident?

\*Does ethnicity play a role in either further susceptibility or resilience to cancer (more or less cancer deaths for certain ethnic groups)? If there is a difference between trends for certain ethnicities in the overall disease burden why would that be the case? (E.G. Presence of a good support system or lack thereof, high religious affiliation, healthier overall living, etc.)

\*Does the particular cancer site have an affect on the outcome of the disease? If so, what sites have the largest mortality relative to the number of cases?

These questions will serve as valuable starting points that, if possible, will allow for deeper data analysis to determine synergy between variables (do certain races or ethnicities have a higher propensity for cancer in certain sites? Are there trends between sex and age group relative to cancer mortality trends?). While some of these questions posed may not be able to be fully explored and answered with the data set utilized here, further research in topics of interest will elucidate these questions and allow for a greater reinforcement of already established research and data analysis regarding cancer and its disease burden.

# 3 Methods and Results

*In most research papers, results and methods are separate. You can combine them here if you find it easier. You are also welcome to structure things such that those are separate sections.*

## 3.1 Data aquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 3.2 Data import and cleaning

library(readxl) #for loading Excel files library(dplyr) #for data processing library(here) #to set paths

#path to data data\_location <- here::here(“data,”“raw\_data,”“United\_States\_Cancer\_Mortality\_1999-2017.xls”)

#load data. rawdata <-read\_xls(data\_location)

#we will now take a look at the data dplyr::glimpse(rawdata)

#We will examine the data more closely to ensure everything is working properly View(rawdata)

Because there are several variables that will not be needed including cancer sites code, age group code, sex code, race code, and ethnicity code, we will filter out the non-desired variables by selecting only the ones we need (the predictors of interest)

The outcome of interest will be Cancer deaths Variables that will serve as predictors of interest will include… sex,age group, race, ethnicity, and cancer sites

#Select only the variables we are interested in to remove the rest processeddata <- rawdata %>% select(“Age Group,” “Sex,” “Deaths,” “Ethnicity,” “Race,” “Cancer Sites”)

#rename Age Group to Age allow for easier data processing colnames(processeddata) <- c(“Age,” “Sex,” “Deaths,” “Ethnicity,” “Race,” “CancerSite”)

#Check to make sure new column name is reflected colnames(processeddata)

#Make sure the processed data looks right dplyr::glimpse(processeddata)

#One last check View(processeddata)

3.2.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.0.1

#Data exploration/description ###################################### #Load packages to aid with data exploration library(ggplot2) #for plotting library(broom) #for cleaning up output from lm() library(here) #for data loading/saving library(tidyverse) #for help with managing data library(scales) #for help with making data more digestible

#Let’s first examine a summary of our data before we continue with exploration mysummary <- summary(mydata) print(mysummary)

#As can be seen from the summary, all variables are considered characters except #for Deaths, which is considered numerical

#Please note…

#Our outcome of interest is Cancer deaths #Variables that will serve as predictors of outcome include… #Age group, sex, total deaths, ethnicity, race, and cancer sites

#Before we do plotting of the data comparing our predictors of interest, we #will examine each variable separately to better understand the data we are #working with

#First we will examine the age group of our data set: mydata %>% ggplot(aes(x = Age)) + geom\_bar() + scale\_x\_discrete(guide = guide\_axis(n.dodge = 1)) + theme(text = element\_text(size = 10), axis.text.x = element\_text(angle = 90, hjust = 1))

#As can be seen from the chart the population is largely composed of older #adults between the ages of 50-85

#Now the sex of the data set: mydata %>% ggplot(aes(x = Sex)) + geom\_bar()

#The chart shows that this data set is almost exactly equal in terms of number # of male and female

#Next we will examine the race of the data set: mydata %>% ggplot(aes(x = Race)) + geom\_bar() + scale\_x\_discrete(guide = guide\_axis(n.dodge = 2))

#As can be seen from the data, we are working with a majority White population #with a moderate amount of African American and a lower amount of Asian or #Pacific Islander and American Indian or Alaska Native

#Now ethnicity: mydata %>% ggplot(aes(x = Ethnicity)) + geom\_bar()

#This chart shows us that the data is composed of a mostly non-Hispanic #population, with Hispanic as the second most common followed by some unknown or #missing

#Finally cancer sites: mydata %>% ggplot(aes(x = CancerSite)) + geom\_bar() + scale\_x\_discrete(guide = guide\_axis(n.dodge = 1)) + theme(text = element\_text(size = 8, face = “bold”), axis.text.x = element\_text(angle = 90, hjust = 1))

#It seems as though the largest number of cancer cases are localized to the #digestive system, lung and bronchus, and respiratory system, with a large number #of cases considered to be miscellaneous

#Now that we have completed a preliminary look at our predictors of interest, #we will examine a closer look at the data to see if any basic trends #can be seen, with the first plot, we will be examining cancer deaths and #race to see if there exists any trends

#Plot 1, Plotting Cancer deaths by Race

Cancer\_Deaths\_By\_Race <- mydata %>% ggplot(aes(x=Race, y=Deaths)) + geom\_bar(stat = “identity”) + ggtitle(“Cancer Deaths by Race in the United States”) + scale\_x\_discrete(guide = guide\_axis(n.dodge = 2, check.overlap = FALSE))+ scale\_y\_continuous(labels = comma)

#Examine plot print(Cancer\_Deaths\_By\_Race)

#As we can see from the chart, cancer mortality is largely concentrated to the #White population group, with African American being the second highest

#Now we will be plotting the number of Cancer deaths by Age Group to examine #If a trend is present between certain age groups having higher levels of #relative mortality

#Plot 2, Plotting Cancer deaths by Age Group

Cancer\_Deaths\_By\_Age\_Group <-mydata %>% ggplot(aes(x=Age, y=Deaths))+ geom\_bar(stat = “identity”) + ggtitle(“Cancer Deaths by Age Group in the United States”)+ geom\_smooth(method=‘lm’)+ scale\_y\_continuous(labels = comma)+ scale\_x\_discrete(guide = guide\_axis(n.dodge = 1, check.overlap = TRUE))+ theme(text = element\_text(size = 12), axis.text.x = element\_text(angle = 90, hjust = 1))

#Examine the Plot print(Cancer\_Deaths\_By\_Age\_Group)

#Like the graph that was viewed earlier, deaths are largely concentrated towards #older populations

# 4 Now we will examine cancer deaths by site to see if there are particular

#sites that are responsible for a higher overall mortality

#Plot 3, Plotting Cancer Deaths by Site

Cancer\_Deaths\_By\_Site <- mydata %>% ggplot(aes(x= CancerSite, y=Deaths)) + geom\_bar(stat = “Identity”) + ggtitle(“Cancer Deaths by Site”)+ geom\_smooth(method=‘lm’)+ scale\_x\_discrete(guide = guide\_axis(n.dodge = 1, check.overlap = TRUE))+ scale\_y\_continuous(labels = comma)+ theme(text = element\_text(size = 6), axis.text.x = element\_text(angle = 90, hjust = 1))

#Examine plot print(Cancer\_Deaths\_By\_Site)

#Once again, as previously discussed in the prior chart examining count of #cancer sites, deaths are concentrated in digestive, lung and bronchial, #and respiratory

## 4.1 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

Example table 4.1 shows a table summarizing a linear model fit.

Table 4.1: Linear model fit table.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| term | estimate | std.error | statistic | p.value |
| (Intercept) | -43.7883068 | 61.1150617 | -0.7164896 | 0.4940713 |
| Height | 0.6996272 | 0.3675692 | 1.9033889 | 0.0934786 |

# 5 Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

This paper (Leek & Peng, 2015) discusses types of analyses.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# References

Leek, J. T., & Peng, R. D. (2015). Statistics. What is the question? *Science (New York, N.Y.)*, *347*, 1314–1315. <https://doi.org/10.1126/science.aaa6146>