**What are the Leading Causes of Death in the United States?**

Wiley Winters

Regis University Anderson College of Business and Computing

MSDS 670 Data Visualization

Mr. John Koenig

March 10, 2024

**Abstract**

This paper describes research on the leading causes of death in the United States from 2000 to 2017. The data was downloaded from the U.S. Department of Health and Human Services and was published by the Centers for Disease Control and Prevention. The scope of this research is to identify the top causes of death and not the underlying social, economic, or behavior factors that may contribute to the death rates.

**What are the Leading Causes of Death in the United States?**

**Research Question**

During the 2020 pandemic, COVID-19 became the third leading cause of death in the United States. The final overall death rate rose from 715 deaths per 100,000 in 2019 to 835 per 100,000 in 2020 due to Covid (CDC, 2022). While COVID-19 caused a lot of deaths, it still did not become the number one or two cause of death in the United States. The research conducted for this paper will answer the question: “What are the leading causes of death in the United States?”

**Data**

All major data sets were obtained from the U.S. Centers for Disease Control and Prevention (CDC) and the U.S. Census Bureau. In addition, a state spelling to abbreviation dataset was manually constructed from resources found on the internet. Data was downloaded as either a CSV or xlsx formatted file. If the file was xlsx formatted, it was converted to CSV to reduce its size and make it easier to manually edit, if required. The author if this paper had difficulty finding continuous U.S. State population statistics on the Census Bureau’s web site. Two CSV files were downloaded. One covers the years from 2000 to 2010 and another from 2011 to 2020. The last two years of census data was not used in this analysis. The Census Bureau files were cleaned up and prepared for analysis. The State population data was required for calculating the *Crude Death Rate* for each U.S. State.

The main data set for this study was downloaded from the CDC’s website in xlsx format and converted to a CSV. This made it easier to manually edit and work with the file if required. The original file consisted of seven columns and 137,700 rows. There are a number of blank or *NaN* values in the dataset, but these were not imputed or removed. According to the data set’s web site, blank values do not indicate the data is missing, it just has not been reported and recorded at time of publication and will be added later. For this analysis, missing values can be tolerated.

**How Death Rates are Measured**

There are three measures used when reporting death rates in literature. The first is the raw number which is just a count of the number of deaths and their causes. The next is age-adjusted rate computed using a direct method by applying age-specific rates in a population of interest to a standardized age distribution. Age-adjusted rates are calculated as in:

where

ri = rate in age group *i* in the population of interest

pi = standard population in age group *i*

P= ∑ni = 1pi

n = total number of age groups over the age range of the age-adjusted rate (CDC, 2022).

The CDC recommends using the age-adjusted rate when comparing populations of different geographic areas. However, it does warn that age-adjusted rates should be viewed as relative indexes rather than actual measures of risk (CDC, 2022). The final measure used in reporting death rates is the Crude Death Rate. These are useful when a person wants to map the observed state or country wide. (CDC, 2018). The formula for calculating the crude death rate is Number of deaths / population \* 100,000. This will give the crude death rate per 100,000 people (Spears, 2024).

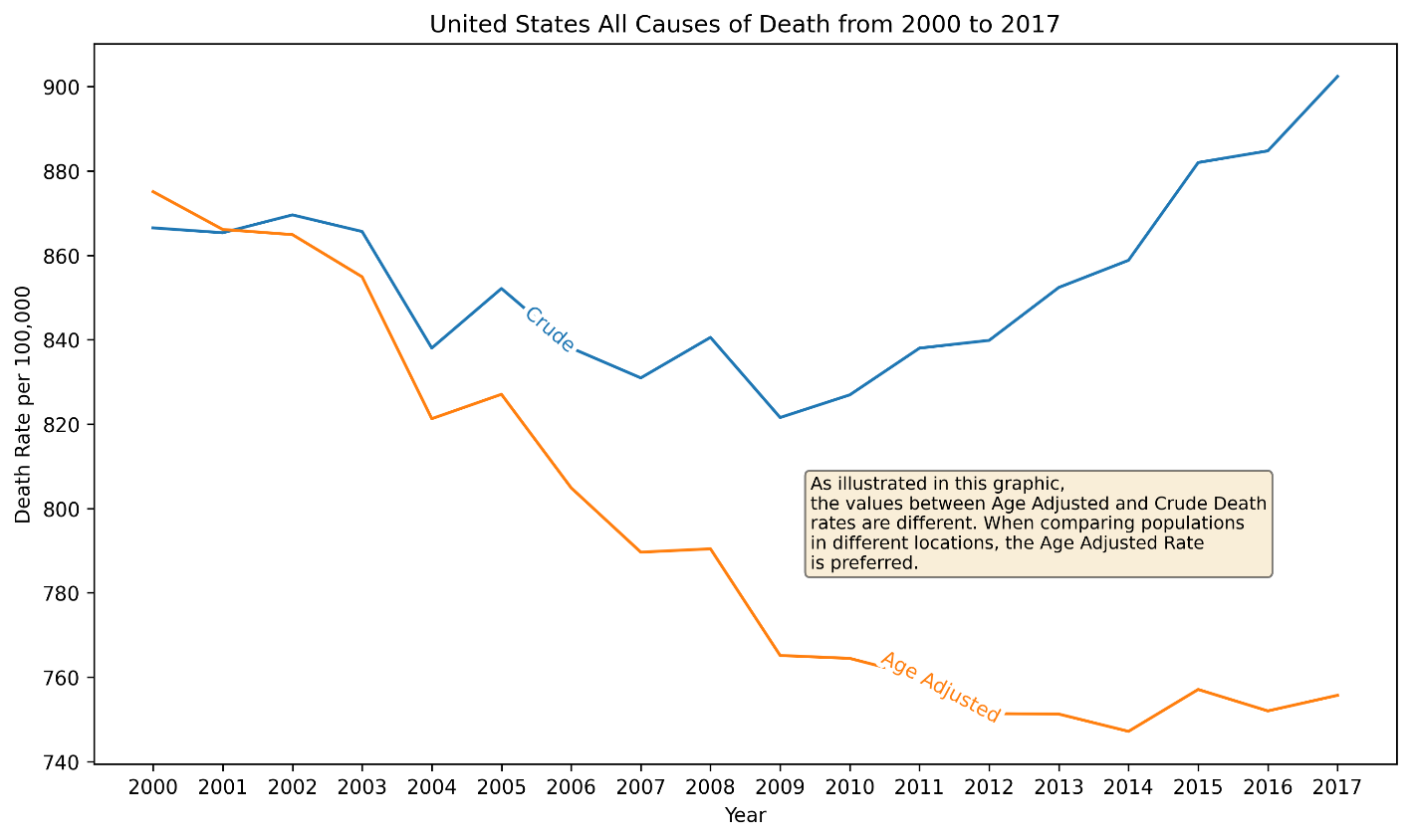
While the crude death and age-adjusted rates are often used in reporting mortality rates, they are not the same. The plot below illustrates how they differ when comparing all deaths in the United States. The age-adjusted rate is lower than the crude rate. 

Figure - Differences between Age Adjusted and Crude Death Rates

**Method**

A review of relevant literature was conducted, and sources were selected for this analysis. Early in the process the author decided to use datasets from trusted sources such as the CDC and U.S. Census Bureau. Both organizations provide access to their data sets and the quality of them is sufficient for conducting analysis and reporting. After the datasets were selected and downloaded, they were manually evaluated using a spreadsheet application, converted to CSV as required, and loaded into *Pandas* data frames using Jupyter Lab. The analysis required the use of data from the CDC and U.S. Census Bureau. This required the merging of three data frames into one. Merging was accomplished by using the pandas.merge() method. Data frames were left joined on the state field.

After creating the main data frame, the crude death rate was calculated for the whole United States and for each state. In addition, a data set was created with state names and abbreviations. This was also merged into the main data for use in creating a choropleth map with death rates displayed on it. Some basic EDA was conducted to ensure the data frame was sane enough for performing an analysis.

**Findings**

The goal of this research is to identify the top causes of death in the United States. Data was filtered to include only relevant information to determine what the leading causes are. Using the crude death rate for the U.S. it was determined that heart disease is the leading cause of death from 2000 to 2017. The top 10 causes are listed in the figure below.

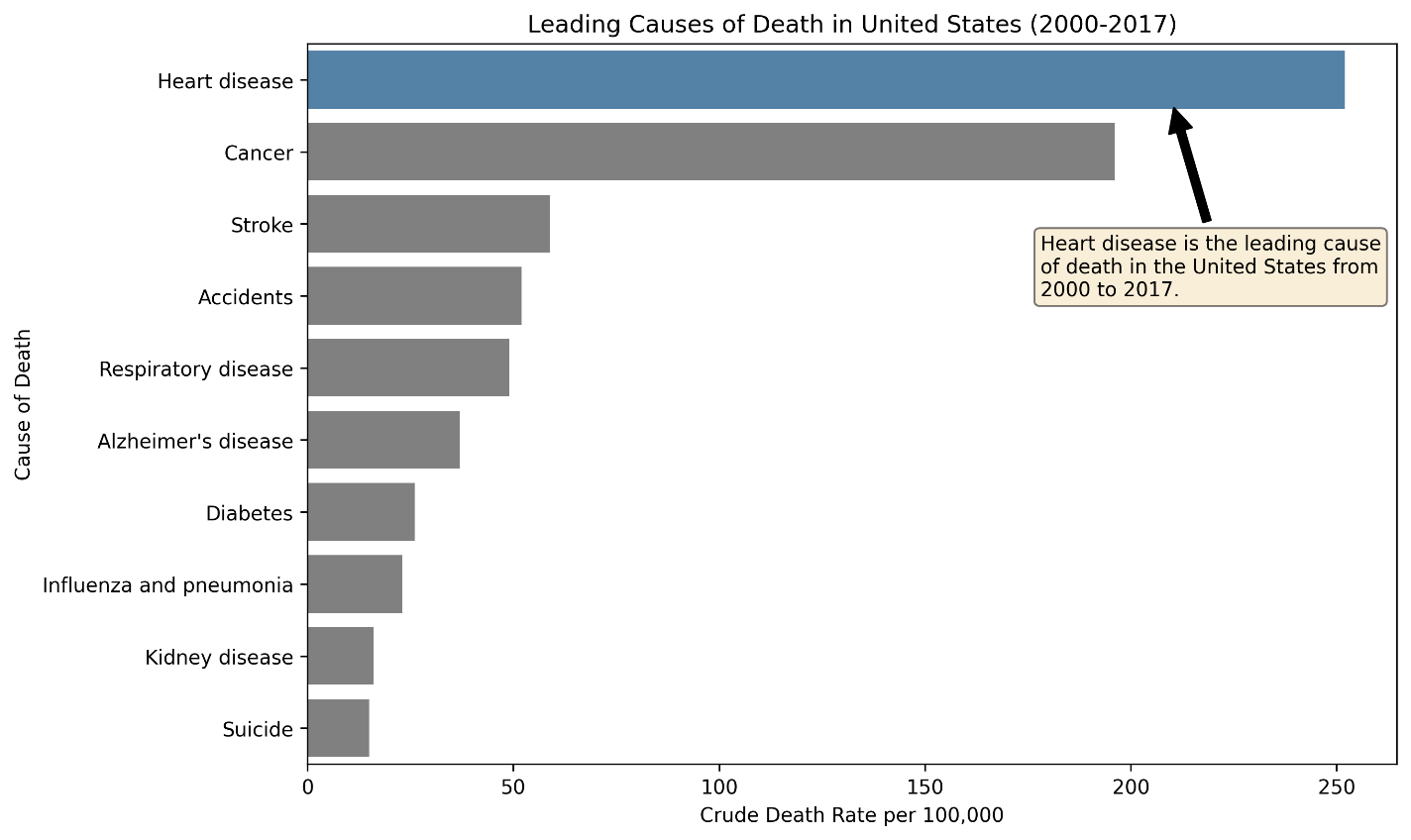


Figure - Leading Causes of Death in the U.S.