Disease Control and Prevention Analysis

Shraddha Kulkarni
shraddhk@usc.edu

Seher Khan
seherkha@usc.edu

December 6, 2019

Abstract

The paper presents an analysis of Mortality rates, Disease Control and Prevention techniques across various States of the United States using advanced visualizations created with the help of d3.js. The main goal of our project is to summarize Mortality Statistics.

Disease Control, Visualization, Analysis, Mortality Rate, Healthcare Expenditure, Clustering

1 Introduction

Some of the leading causes of death in the United States include heart disease, cancer, alcoholism and injuries caused by accidents. Many of these are preventable, and being able to visualize the numbers could help us develop improved strategies to avoid and alleviate their occurrence. We expect our application to be useful for the general public as it will make them more aware of the diseases they are susceptible to based on their location and other demographic traits. Medical students and researchers could also benefit from knowing which fields require their most interest. Finally, policymakers would greatly benefit from a tool that would help identify major health problems in their region, enabling them to draft effective policies. Therefore our project comprises of a detailed study of Historical data on Mortality Rates and Causes of Deaths from the year 1950-2016 where we highlight interesting trends using various visualizations ranging from, for example, simple forms such as bar charts to more complex tools such as interactive bipartite diagrams.

2 Data

2.1 Data Collection

The data employed in this study has been extracted from the Centers for Disease Control and Prevention (CDC) website. CDC is one of the major operating components of the Department of Health and Human Services. This protection agency not only works toward preventing, controlling and eradicating disease, it also serves as a hub for health related statistics. This includes highly faceted data tables for topics such as specific diseases/injuries, births and abortions and costs associated with healthcare. These datasets are provided in a variety of forms including tables, charts, presentations and reports. Due to the scope of our project, and also limitations of time and resources, we have only utilized a subset of this wealth of information — focusing on,

- fatality statistics by race, sex, gender, age, cause of death and location, over time; and
- the expenditure on health care and the sources that are used to pay for it.

.

2.2 Data Preprocessing

The Pandas library in Python was employed to preprocess the data for each visualization. For example, it was used to read data of multiple years, group them to find average values and store them as a CSV file, or used to dynamically create a JSON file. The Filter, Sort and Pivot Table features of Google Sheets and Microsoft Excel was also used for preprocessing data.

3 Approach

3.1 Design Consideration

Our design exhibits an equal balance between unidimensionality and multidimensionality. Various factors such as gender, race, origin, age-group, and expenditure are displayed using Maps, graphs, and charts. Interactivity ensures multidimensionality. For our audience's understanding, we have used familiar charts and avoided redundancy. The visualization is more abstract than figurative because most of the references applied are conceptual tools such as bar, line, and pie rather than vivid physical realities. Sticking to functionality has been our forte.

3.2 Technical Considerations and Development Plan

Our visualizations were created using D3.js which is a JavaScript library for manipulating documents based on data. We also used node modules such as topojson, particles.js and Bootstrap (particularly for CSS styling, creating buttons and for the grid layout) to improve functionality and make the website visually appealing.

The web application was developed on the Vue.js framework. Each page of the website is created using a Vue View. In case of pages with only one visualization, the script for the visualization was directly mounted on the View. In case of pages with multiple visualization, the View held multiple Vue Components instead — each component receives data from the parent View and generates the graph, which is then inserted in the View. Finally, the system was finally deployed on pdms.usc.edu which is a nginx web server using FileZilla.

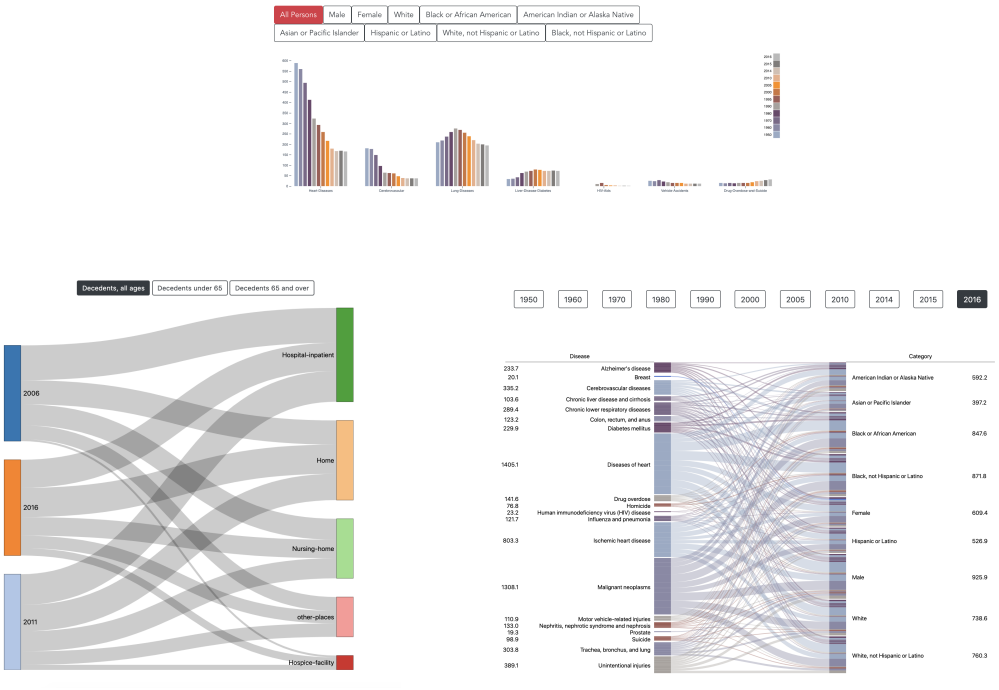
4 System

4.1 Disease-Category Correspondence

In order to identify concerning categories with respect to various diseases, we constructed a Bipartile Graph and Grouped Bar Chart for our first visualization. Both of these present the death rates for the selected years from 1950-2016 but they serve different purposes.

The goal of the Bar Chart is to highlight the trend of the death rates of particular disease for a given segment of the population. Hence, data points have been grouped on the basis of types of diseases (such heart diseases and lung diseases) such that their trends over the period can be assessed side by side. Selector can be used to filter for a particular gender, race, and place of origin in the chart. So for example, selecting African Americans, one can see that the death rates have reduced over the years due to heart diseases, whereas there has been a fluctuating trend for death rates of lung diseases.

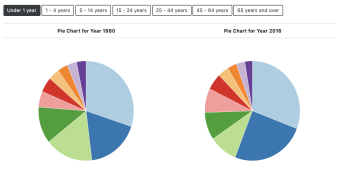
On the other hand, the Bipartite graph serves to inform the user how strongly the diseases contributed to deaths of a segment of the population and for a particular year. We do this by presenting a many to many correspondence between diseases and categories (i.e. facets of the population such as male, female or Asian). Note that this a two way chart and so we can explore the breakdown from both sides: selecting a disease from the left side will give us a breakdown of deaths due to that disease. Selecting a category from the right side displays a breakdown of diseases that have led to deaths for a particular community of people. Finally, selector buttons are provided to set the year. The reason for avoiding using separate graphs for groups such as male/female, and separate charts for races is two-fold: the data did not permit multi-faceting (such as all Caucasian males), and because it is visually obvious to the user that, for example, data for males should be compared to the date for females.



We have also made a Bipartite Graph which displays one-to-one correspondence between the Diseases and Category. This figure helps us to analyze the major reasons for Death for particular sex, race or origin.

Grouped Bar chart helps us to analyze the trends over the years for a particular category. Selecting sex, race, or origin, the screen dynamically displays a vertical bar chart for major Diseases and their height is the Death rate over the years. It starts in 1950 and goes on until 2016, with a gap of 5 years. The major diseases covered included Hearth Diseases, Cerebrovascular, Lung and Liver Diseases, Diabetes, HIV Aids, Vehicle Accidents, Drug Overdose and Suicide. The Heart and Cerebrovascular diseases have shown a steady dip, whereas Lung and Liver Diseases have a fluctuating trend. There has been an increase in deaths due to drug overdose and Suicide, which is a clear indicator of the dire Mental health conditions across the globe.

4.2 Age Group Mortality Exploration - I



Category	Rank	Color	Causes of Death in 1980	No. of Deaths in 1980	Causes of Deaths in 2016	No. of Deaths in 2016
Under 1 year	1		Congenital anomalies	9220	Congenital malformations, deformations and chromosomal abnormalities	4816
	2		Sudden infant death syndrome	5610	Disorders related to short gestation and low birthweight	3927
	3		Respiratory distress syndrome	4089	Sudden infant death syndrome	1500
	4		Disorders relating to short gestation and unspecified low birthweight	3668	Newborn affected by maternal complications of pregnancy	1402
	5		Newborn affected by maternal complications of pregnancy	1572	Unintentional injuries	1219
	6		Intrauterine hypoxia and birth asphyxia	1497	Newborn affected by complications of placenta, cord, and membranes	841
	7		Unintentional injuries	1166	Bacterial sepsis of newborn	583
	8		Birth trauma	1058	Respiratory distress of newborn	488
	9		Pneumonia and influenza	1012	Diseases of circulatory system	460
	10		Newborn affected by complications of placenta, cord, and membranes	985	Neonatal hemorrhage	398

The visualization consists of two pie charts and a table, for the years 1980 and 2016. There has been an attempt to compare the leading causes of death in the two respective years. Considering there are more than 35 years of difference, there is an expectation that the causes which were more relevant back then have taken been brought under control. Unintentional injuries, Congenital malformations, chromosomal abnormalities, and Malignant neoplasms are major reasons for death for kids aged 4 years or less. Whereas, Suicide and Homicide seem to be more significant for

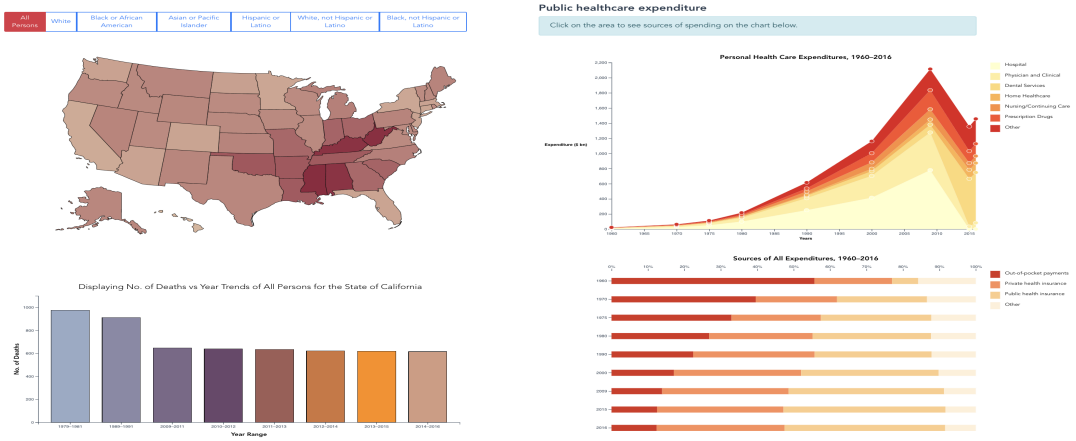
children between the age of 5-24 years. While 40+ people mostly have suffered from Heart Diseases, Diabetes, Liver and Respiratory Diseases.

4.3 Age Group Mortality Exploration - II

In order to visualize death rate by age groups we used a HeatMap and a Grouped/Stacked Bar Chart. To explore the distribution of the data we have used a Heatmap. Each cell is labelled by death rate of that particular age group and year. The default "overall" view colors each cell according to it's number, irrespective of the year or age group. This allows the user to immediately identify an outlier/concerning datapoint. In our case, no such outliers existed. The user can then either select "Age" or "Years", through which the cells color range can be altered. For example, by selecting "Year", the color scheme is based only on the death rates of each year, and we can see that distribution of deaths across age groups has remained roughly the same over the decade. On the other hand, by selecting "Age", the color scheme is based only on the death rates of each age groups, and so we can observe that only in the 1-14 and 65+ years categories there is a clear and consistent decline in deaths.

The Stacked/Grouped bar chart provides the same data as the Heatmap but with greater granularity, allowing the user to spot more subtle trends. It also allows improved comparision between groups by providing the functionality to hide/unhide series.

4.4 Geographic Mortality Exploration and Healthcare Expenditure



This visualization consists of a Choropleth and a Bar Chart. It represents the data from the year 1970 to 2016. Clicking a particular race, or origin, the choropleth changes dynamically. The page also enables the user to click on particular States to look at the individual Death trends over the years in a particular State.

The Stacked Area Chart depicts the trend in public health expenditure in the US by category since the 1960s. We can see that health care expenditure shot up since 1980 from a little over 200bn to 2100bn dollars in 2010, since then it has fallen and stabilized around 1400bn dollars. In the future, it would be interested to see if this is a result of large population growth or other factors.

The complete Stacked Bar Chart shows the source of these expenditures over the same period. In default view (which is for total healthcare expenditure) we can see that over half of the spending was sourced from out of pocket payments in 1960. However, this rapidly declined and was replaced by both public and private insurance.Clicking on a specific category in the area chart (or the legend), update the second chart. This allows one to see trends in the source of funding for a particular category. Clicking on the white chart area resets the second chart.

4.5 Clustering Diseases by Region

We finally wanted to highlight the most concerning diseases in the US in the past few years. Over 2.7 million fatalities occurred in the US during 1999-2016 and only 10 diseases/causes contributed to 3 out of 4 fatalities. To display the relative values of these 10 disease/causes amongst each other we created a simple Bar Chart where the height of the bars are proportional to the number of fatalities caused by that disease/cause. This chart highlights that the heart diseases and malignant neoplasms, by far, contribute to the greatest count of deaths in the country.

Since users are often considered with a specific region or state, we also created a Circle Packing Chart that showed the top 5 diseases of each state in the US. Note that we found that while kidney disorders (nephritis/nephrotic syndrome) was 9^{th} most common cause of death in the US, it was not among the top 5 diseases of any state; on the other hand, the remaining 9 most fatal diseases were. To allow it to be easy for users to focus on one or a few causes of death, and compare their relative size, we added the ability to highlight their corresponding bubbles with specific colors. This shows us, for example, that heart disease is and malignant neoplasms are the most common fatal diseases in each state.

5 Conclusion

The main goal of our project summarizes Mortality Statistics. We aimed at creating visualizations that will be useful for anyone who wants to get insights about the Health data. We wanted to create a site that is user-friendly, interactive, and knowledgeable. There were more than 100 datasets on the CDC website, deciding on the most important information, merging and preprocessing the datasets was a huge part of our project. We also particularly focused on creating an aesthetically pleasing site. The fonts and colors have been used accordingly. Overall, a lot of thought has gone over this project, and we believe it will be helpful for the general audience.

6 References

1. Michael Bostock, Vadim Ogievetsky and Jeffrey Heer.: D3: Data-Driven Documents.
4. Murray, S.: Interactive Data Visualization for the Web.O'Reilly Media (2013)
5. Ware, C.: Visual Thinking for Design. Morgan Kaufmann (2008)
6. D3.js, https://d3js.org/
7. The National Center for Health Statistics. 2018.
Available from: https://www.cdc.gov/nchs/products/nvsr.htm