final

Jen Johnson

12/5/2017

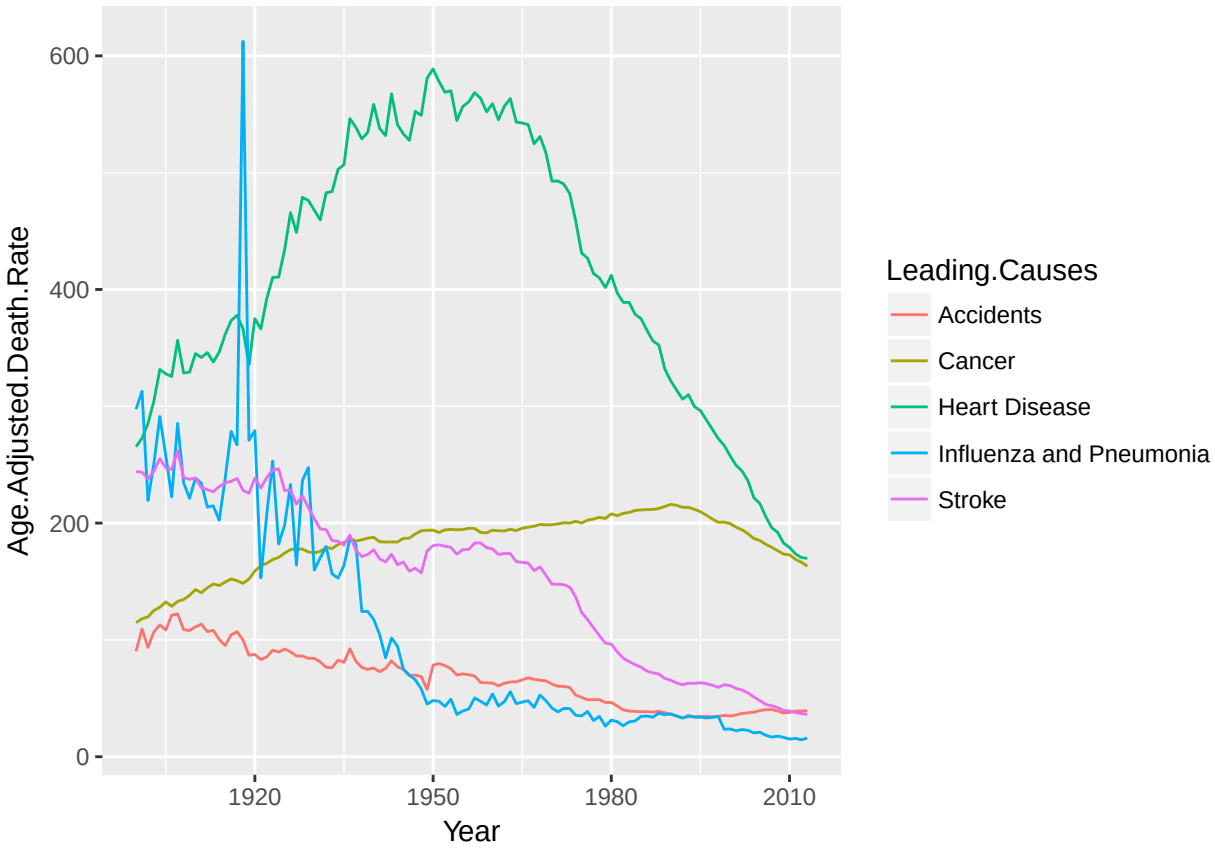
To add to Introduction

Data.gov and Project Tycho disease datasets will be used to visualize how the prevalence of disease-caused deaths has been impacted by vaccines. It is hypothesized that the effects of vaccines will be seen in the data. It is hoped that both the short term and long term effects of these drugs will provide insights on the efficiency of such developments and their prospects for the future. The data will be represented in multiple ways to demonstrate its different stories and interpretations.

Motivation

The first dataset originally came from a data.gov source, but was downloaded from kaggle. It shows the leading causes of death from 1900 to 2013. It is a subset of a dataset used later in the project, and only includes 5 broad categories.

```
data <- read.csv("Death_Rates1900-2013.csv")

ggplot(data, aes(x = Year, y = Age.Adjusted.Death.Rate)) + geom_freqpoly(aes(color = Leading.Causes), s
```



The x axis is time, and the y axis is age adjusted death rate. This means that causes of death of different demographic groups are scaled for easier and more accurate comparisons. The colors are the different categories of causes. We can see that there are definite differences in trends for the different causes of death.

The light blue influenza and pneumonia line peaks in 1920, as would be expected because of the WWI outbreak. Then, it drops. This could be because the discovery of penicillin in 1928 by Alexander Fleming. The decrease is steady until 1950. The period between 1920 and 1950 has been called the antibiotic era. However, because of the evolution of resistant species, the efficiency of antibiotics and vaccines has not decreased as dramatically in recent years as during this period.

These results provided inspiration for the rest of this section, whose focus is contagious diseases. We have seen a decreasing trend in influenza and pneumonia because of developments in medicine. How have the frequencies of diseases other than influenza and pneumonia changed over time?

Methods

How has the prevalence of disease-caused death changed over time?

The datasets obtained were part of Project Tycho. The goal of the project was to increase the availability of public health data. There is information for the diseases hepatitis A, measles, mumps, whooping cough, polio, and rubella. Each dataset includes weekly data per state between 1916 and 2010. The two measures of prevalence are number of cases and incidence rates scaled by population of the state. I decided to use incidence rates. I grouped the data into year, because the vast volume of data was too much to map and the fine-grained week view did not contribute to the user's understanding of changes in disease prevalence.

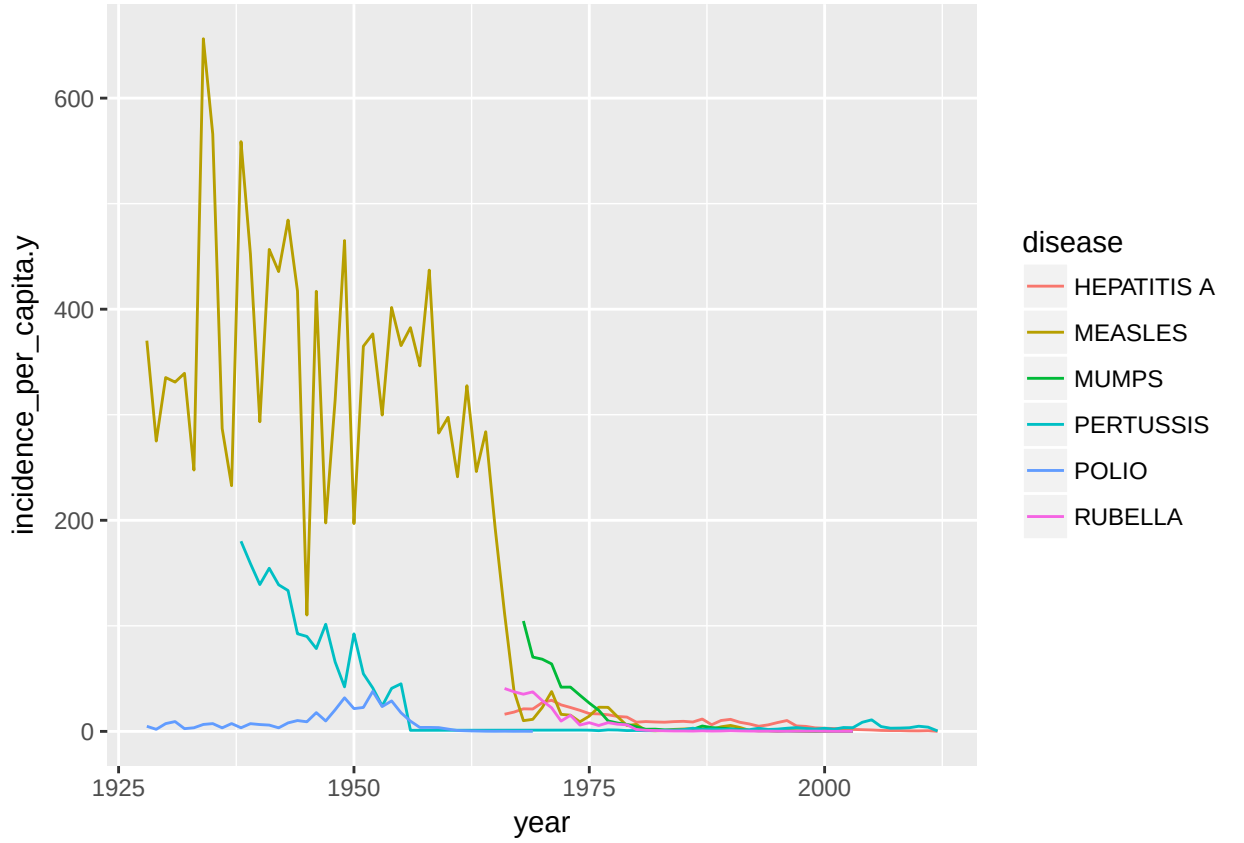
I also combined longitude and latitude data for each state with the prevalence data to be able to map the data points in a logical fashion using the leaflet package.

```
# read in data
disease.directory <- "contagious-diseases"
disease.files <- list.files(disease.directory)
all.diseases <- data.frame()

for (i in 1:length(disease.files)){
    #print(disease.files[i])
    current.data <- read.csv(paste(disease.directory, disease.files[i], sep = "/"))
    all.diseases <- rbind(all.diseases, current.data)
}

# convert into sensible year format
all.diseases <- all.diseases %>%
  mutate(year = round(week/100))

# aggregate state + week data
all.diseases$incidence_per_capita <- as.numeric(as.character(all.diseases$incidence_per_capita))

temp <- aggregate(incidence_per_capita ~ year + state + disease, all.diseases, sum)

all.diseases <- merge(all.diseases, temp, by = c("year", "state", "disease"))

all.diseases <- all.diseases %>% select(year, state, disease, state_name, incidence_per_capita.y)
```

This plot shows the mean incidence rates for each disease over time. The state and week data are aggregated.

```
j <- aggregate(incidence_per_capita.y ~ year + disease, all.diseases, mean)
```

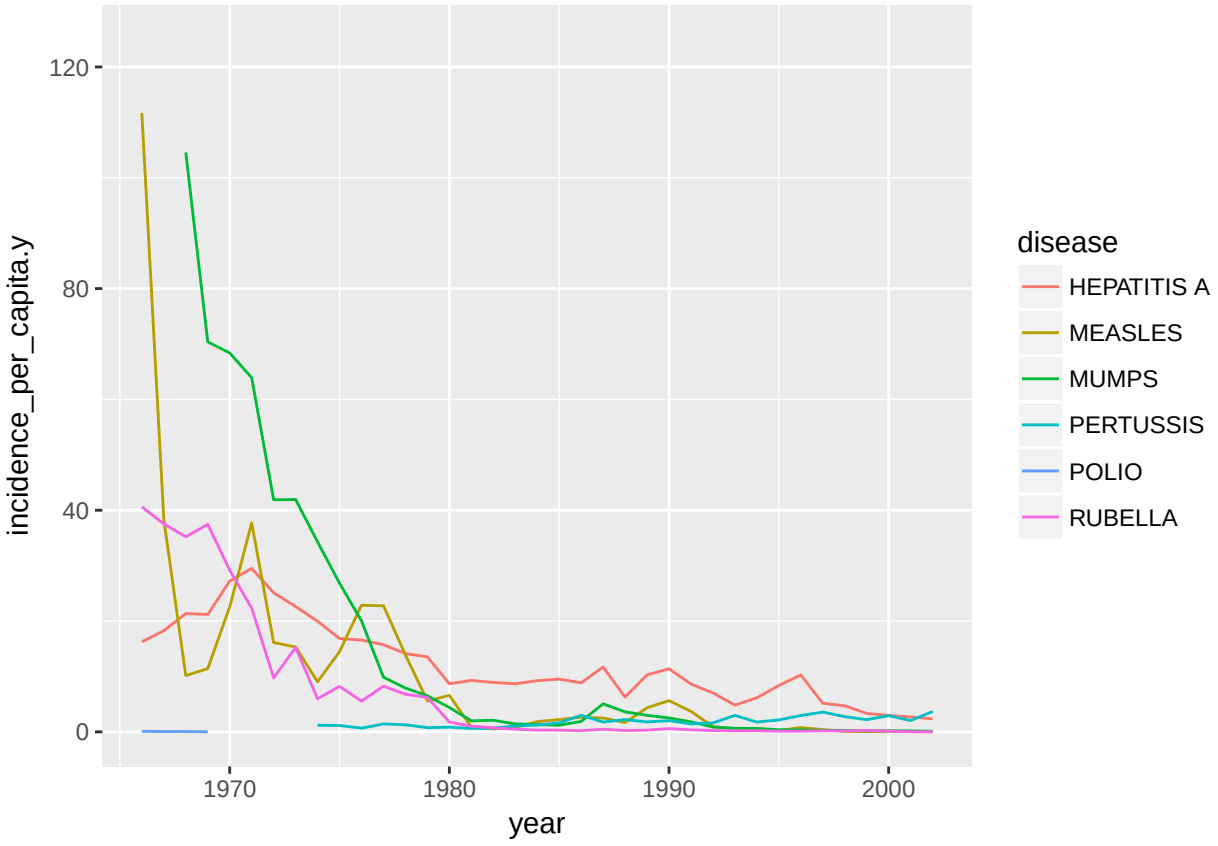
```
ggplot(j) + geom_line(aes(x = year, y = incidence_per_capita.y, color = disease))
```



The fluctuation in the yellow-brown line for measles shows that disease prevalence varies greatly. However, there is a general trend towards fewer cases in recent times.

```
ggplot(j) +
  geom_line(aes(x = year, y = incidence_per_capita.y, color = disease)) +
  ylim(0, 125) +
  xlim(1966, 2002)
```

```
## Warning: Removed 118 rows containing missing values (geom_path).
```



With a fine-grained focus, we can see variation in more of the incidence per capita diseases. This range of years, 1966-2002, overlapped most with the majority of the data so this range was used in the Shiny app.

Results

Click here to open the visualization.

The two sides of the application are the same, and are used to observe two time periods at once. The left side can be animated over time. Any number of diseases can be selected. When errors occur, it means that (one of) the disease(s) selected does not have data for the time point selected. Use the previous figure (with x axis between 1970 and 2002) for reference of missing data. The radius of the dots or the height of the bars are the incidences per capita for each state. They can be clicked to see the label for incidences per capita.

The first example where the impacts of vaccines can be seen in the data is the Measles, Mumps, and Rubella vaccine. It was developed in 1971. In 1970, the prevalencies of these 3 diseases are high and generally equal for most states. As time goes on, the height of the bars that represent incidence per capita decreases. The overall trend shows that the vaccine was effective at reducing the number of cases. However, even in 2002, the most recent date included in the dataset, the height of the mumps bars are not as low as the other diseases.

The app can also be used to compare 1 disease at a time, since most vaccines are species specific. The hepatitis vaccine was developed in 1971, so two years to compare could be 1970 and 1975. There are no more larger bubbles in 1975. However, as you continue along the timeline until 2000, larger bubbles start to appear again. The largest bubble represents a value of 11, which is much less than in 1975. However, the radius is misleading unless you read the label.

Therefore, a different approach was tested to facilitate visualization over time. These maps were created using the same dataset. A heatmap was used instead of radius to represent the incidence per capita.

```
# merge with a different lat/long source to produce a polygon for each state
state_map <- map_data('state')
```

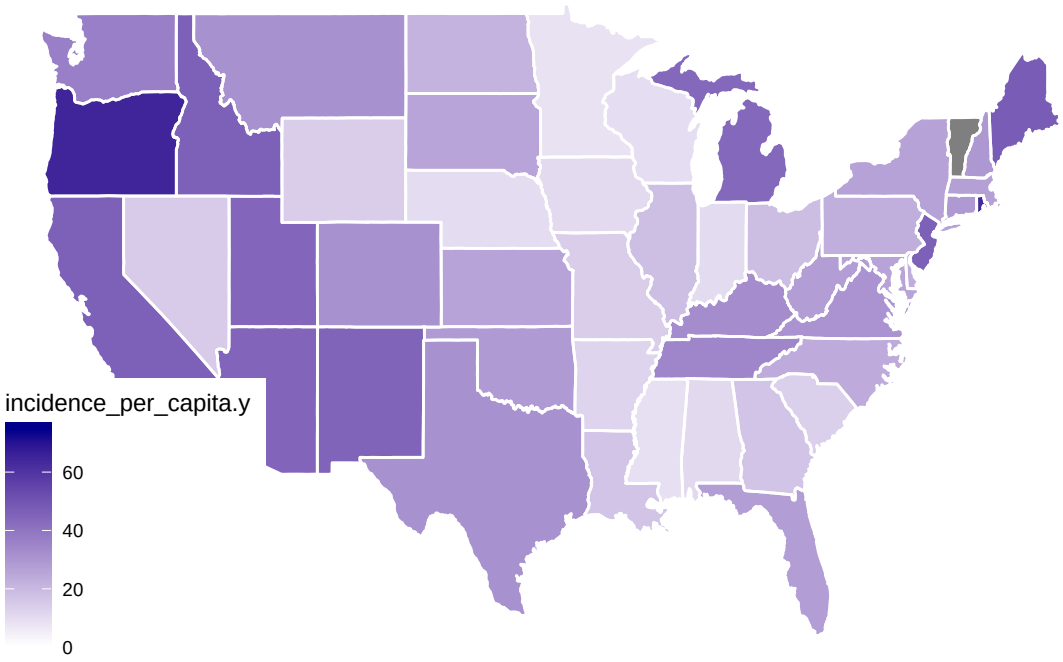
```
##
## Attaching package: 'maps'
```

```
## The following object is masked from 'package:purrr':
##
##     map
```

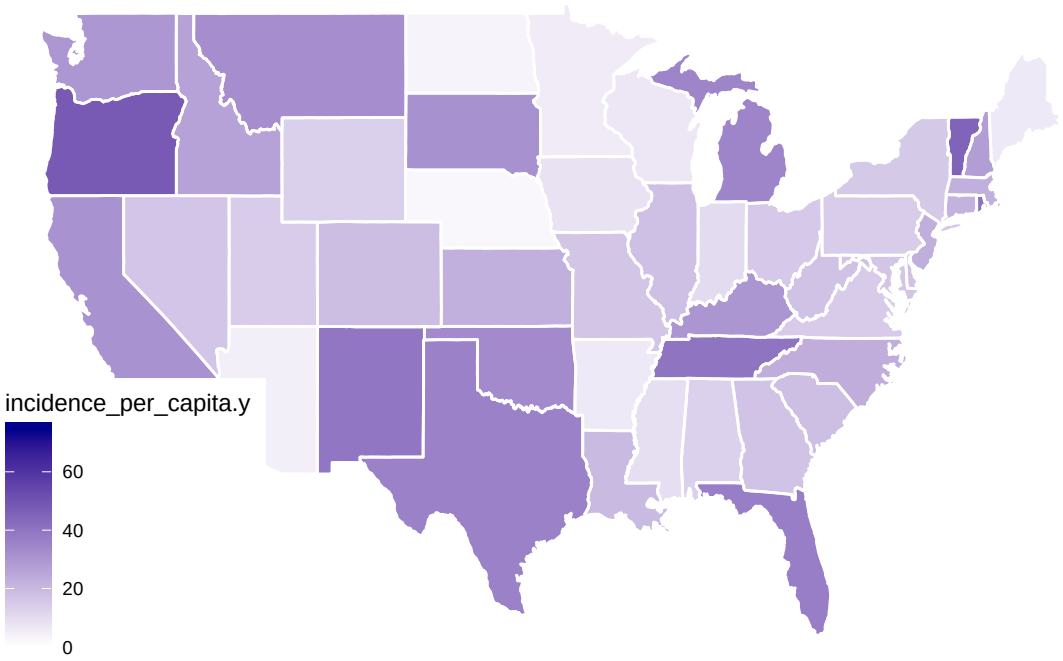
```
all.diseases$region <- tolower(all.diseases$state_name)

single.diseases <- all.diseases %>%
  group_by(region) %>%
  right_join(state_map, by = "region")
```

```
#Plotting heat maps for the Heptatitis in 1971, 1973, 1975, and 2000.
d <- c("HEPATITIS A")
years <- c(1971, 1973, 1975, 2000)

for(i in 1:length(years)){
  current.year <- years[i]

  current <- single.diseases %>%
  filter(disease %in% d) %>%
  filter(year == current.year)

  p <- ggplot(current, aes(x = long, y = lat, group = group, fill = incidence_per_capita.y)) +
  geom_polygon() +
  geom_path(color = "white") +
  theme_map() +
  scale_fill_gradient2(low = "white", high = "darkblue",
                       limits = c(0, 75)) +
  ggtitle(paste("Prevalence of Hepatitis in", current.year))

  print(p)
}
```

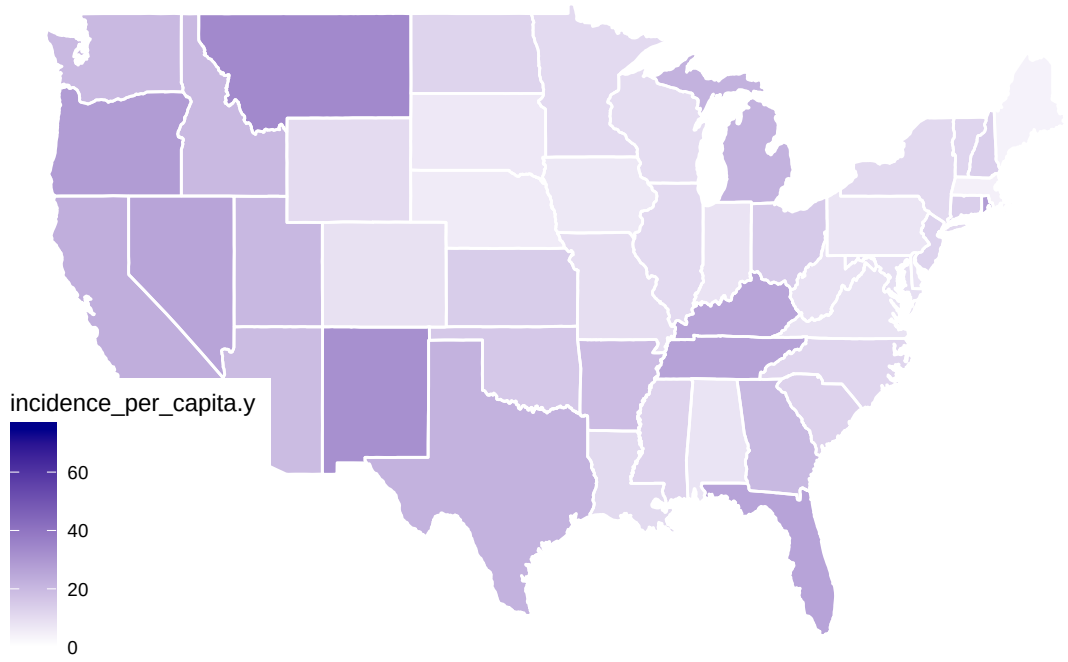
Prevalence of Hepatitis in 1971



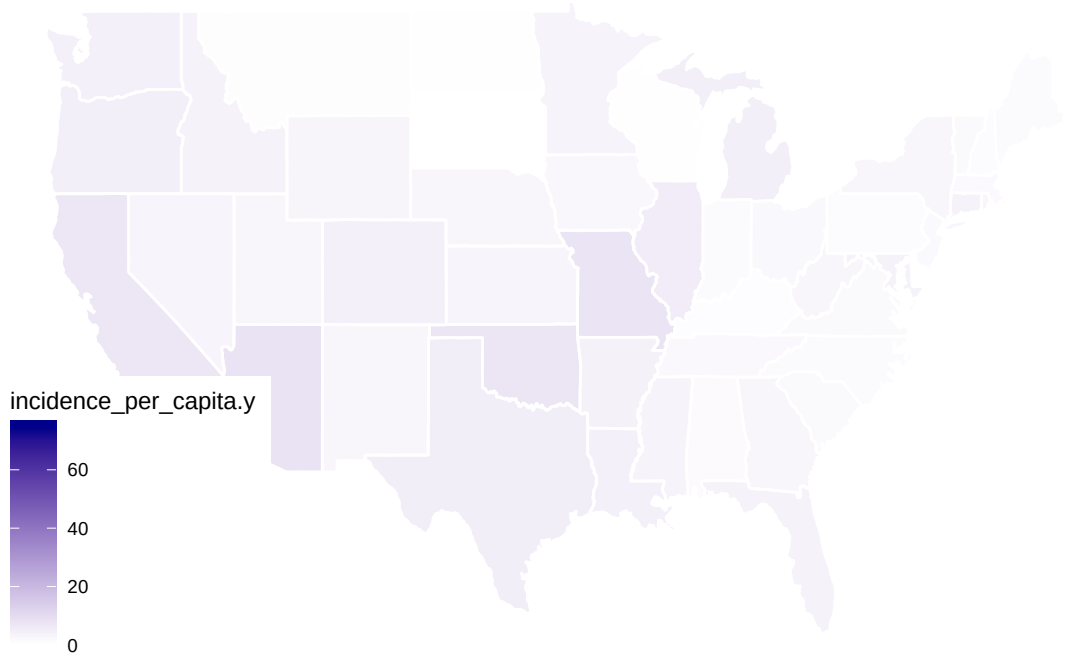
Prevalence of Hepatitis in 1973



Prevalence of Hepatitis in 1975



Prevalence of Hepatitis in 2000



All four maps have the same scale from 0 to 75. The darker the shade of purple, the higher the incidence rates. A grey state means that data is missing. The later 2 maps are overall lighter shades of purple. It is easier to see the effect of the hepatitis vaccine in 1972 because the colors are easier to differentiate than differences in radii. Furthermore, the scale is normalized over time instead of scaled to distinguish betweeen states. However, this method was not used in the app because the data processing takes too long.

The effects of vaccines on dieseases and the deaths they cause have been visualized in a variety of ways. Modern developments in healthcare have been effective, but are not the ultimate solution. Bacterial evolution of antibiotic resistant genes means that the evolution of vaccines (ie their continued development) is essential.

To add to Conclusion

It has been seen that the although vaccines and antibiotics have been effective at reducing the prevalence of disease, deaths caused by contagious disease are by no means eliminated in modern times. Deaths caused by disease still fluctuate. However, the smaller scale of these fluctuations provides evidence for successful developments in medicine.

Sources

"Leading Causes of Death in the USA". Scraped from data.gov by LiamLarson. Kaggle. https://www.kaggle.com/kingburrito666/leading-causes-of-death-usa/data.

"Project Tycho: Contagious Diseases." Compiled by a team at the University of Pittsburgh. Kaggle. https://www.kaggle.com/pitt/contagious-diseases.

"USA lat, long for state abbreviations." Washim Ahmed. Kaggle. https://www.kaggle.com/washimahmed/usa-latlong-for-state-abbreviations.

BIOL 310: Microbiology. Middlebury College. F17. For inspiration and information about vaccines.