Contagious Diseases in the United States: Trends and Cycles in the Past 100 Years

Chen Song
SCI, University of Pittsburgh
135 N Bellefield Ave.
Pittsburgh, PA
chs222@pitt.edu

Jiexiao He
SCI, University of Pittsburgh
135 N Bellefield Ave.
Pittsburgh, PA
jih102@pitt.edu

Jingran Xie
SCI, University of Pittsburgh
135 N Bellefield Ave.
Pittsburgh, PA
jix73@pitt.edu

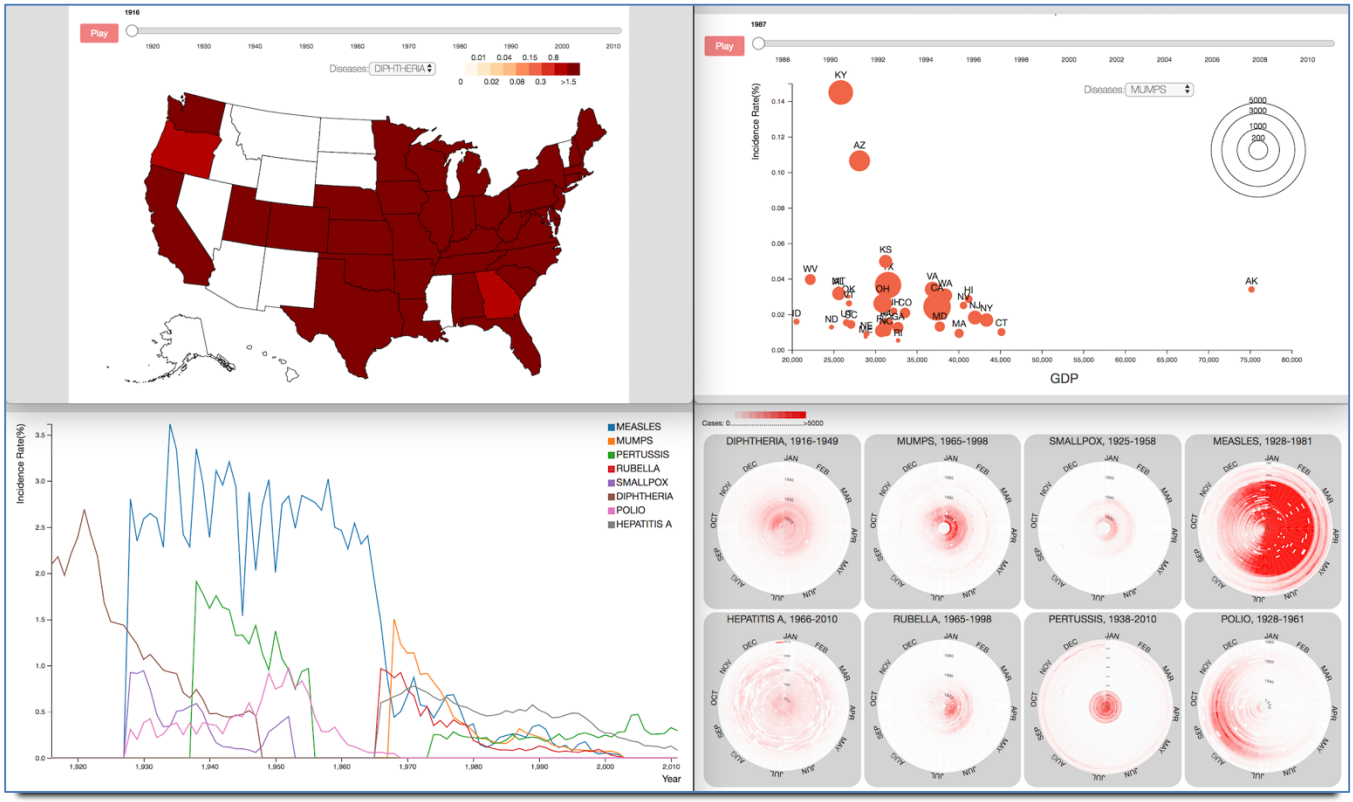

Figure 1: *An overview of the visualizations*

ABSTRACT

Project Tycho provides great datasets which contain weekly counts of cases or deaths of more than 50 contagious diseases at state or city level around the US for more than 100 years. With these datasets, it will be easier to explore the epidemic spreads and preventions in the US from a historical view. However, due to the relatively large data size and the spatial and temporal range of the datasets, it will be hard to reveal any trends or patterns through scanning the long table. Data visualization could provide useful tools and views for us to explore the datasets. In this project, we will design a series of visualizations to reveal the long-term trends and patterns of contagious diseases in the US.

KEYWORDS

Information visualization, Project Tycho, Contagious Diseases

1 INTRODUCTION

In the history of medicine and the history of human, contagious disease is always an important issue for all countries and people. In the history, the "Black death" is estimated to have killed 30–60% of Europe's total population. (Wikipedia: Black Death, https://en.wikipedia.org/wiki/Black_Death). Most recently, SARS and Zika brought great terror to lots of people. With the development of medicine, some diseases like Smallpox were finally defeated by vaccinations, while some others are still hard problems for public health, i.e. HIV. Therefore, it is very important for us, both from a medical aspect and a more general social science view, to study the history and current situation of contagious diseases and get a clearer understanding about the important issues in contagious diseases' spreads and preventions.

Project Tycho provides great datasets to support the study on this topic, which contains weekly counts of cases or deaths of

more than 50 contagious diseases at state or city level around the US for more than 100 years. However, due to the relatively large data size and the spatial and temporal range of the datasets, it will be hard to reveal any trends or patterns through scanning the long table. In this project, we will design a series of visualizations using datasets from Project Tycho, to reveal the three kinds of long term trends and patterns about contagious diseases in the US.

2 RELATED WORKS

There are already many works explored Project Tycho datasets and the history of contagious diseases in the US. For example, Y. Matsubara et al. (2014) defined 5 properties from Project Tycho data: (P1) disease seasonality; (P2) disease reduction effect, mainly due to vaccination; (P3) local/state-level sensitivity; (P4) external shock events, i.e. World War II; (P5) detect incongruous values.

In another research, W. G. Panhuis et al. (2013) indicated that for the eight kinds of vaccine preventable diseases, “declines in the incidence of contagious diseases in the United States over the past century. However, some contagious diseases are now on the rise despite the availability of vaccines. ”

These two papers provide great guides for our work. In our project, we plan to show three kinds of basic trends and patterns in the history of contagious diseases spreads and preventions using information visualization designs: (1) In a long range of time, some of the diseases were significantly reduced by vaccinations while others were not; some were reduced but revived later; (2) Some of these epidemic diseases have significant (seasonal) cycles while others do not; (3) Respecting the geographic locations, some places are more vulnerable to some diseases.

3 DATA DESCRIPTION AND PREPARATION

Project Tycho contains three datasets. Level 1 data contains different types of counts of 8 diseases in 50 states and 122 cities from 1916 to 2010 which have been standardized in a common format. Level 2 data contains informational counts of 50 diseases in 50 states and 1284 cities from 1888 to 2014 which have been reported in a common format. Level 3 data contains different types of counts of 58 diseases and 81 disease subcategories in 3026 cities which have not been standardized. Due to the large size of level 2 and level 3 data, we first choose level 1 data to design and test our visualizations.

The current version (1.0.0) of level 1 data includes counts at the state level for smallpox, polio, measles, mumps, rubella, hepatitis A, and whooping cough, and at the city level for diphtheria. It is actually a subset of level 2 data which was cleaned further and used for a study on the impact of vaccination programs in the US (W. G. Panhuis et al. 2013). In level 1 data there are 7 fields:

- epi_week: the time id of each tuple, in the form of "yyyyww", where "ww" is the "ww"th week of a year, counted from 1 to 52;
- state: the abbreviation of each state;
- loc: the name of state or city;
- loc_type: state or city;
- disease: the name of disease;
- cases: count of cases;
- incidence_per_100000: incidence rates per 100,000 population based on historical population estimates.

The dataset is in a good format. However, in our visualization design, we will use a lot of aggregated data, like the count of each disease per year instead of per week. To avoid real-time calculation as much as possible and speed up the visualization, we preprocessed the data through aggregating it at different levels depending on different needs of each visualization, and convert the original csv file to json files.

4 VISUALIZATION DESIGN

4.1 Trends in the Past 100 Years

In the history of contagious diseases' spreads and preventions, some of the diseases were significantly reduced by vaccinations. This trend could be influenced by a lot of factors from variety sources. The development of medicine and vaccines is definitely the most important factor in contagious diseases' preventions. But at the same time, the general economic development is also an important factor, which could determine the capacity of mass production of safe and effective vaccines, as well as the capacity to distribute the vaccines to majority population. For example, before the development of a modern vaccine for Smallpox, there were a method of inducing immunity known as inoculation used in China from Ming Dynasty. It had a 0.5–2.0% mortality rate, but that was considerably less than the 20–30% mortality rate of the disease itself. (Wikipedia: Smallpox vaccine). However, only a few people could get this treatment due to the underdeveloped economy. Smallpox kept on worrying Chinese people until 1961.

On the other side, some diseases were first prevented by vaccines, but revived later. According W. G. Panhuis et al. (2013), this could be caused by the decrease of risk to get sick from the disease: from the aspect of each individual, when the over-all risk of disease decreased and the risk of vaccine's side effects remains the same, it is reasonable to choose not to get any vaccine, which in general, could cause a decrease of vaccination rate and increase of disease incidence. Also, according to Y. Matsubara et al. (2014), the general trends could be interrupted by some special events, like World War II.

To visualize a time-series data and focus on the whole shape of the data, line chart is definitely our first choice. We design a multiline chart to visualize the trends of all the 8 diseases in the 100 years as follow:

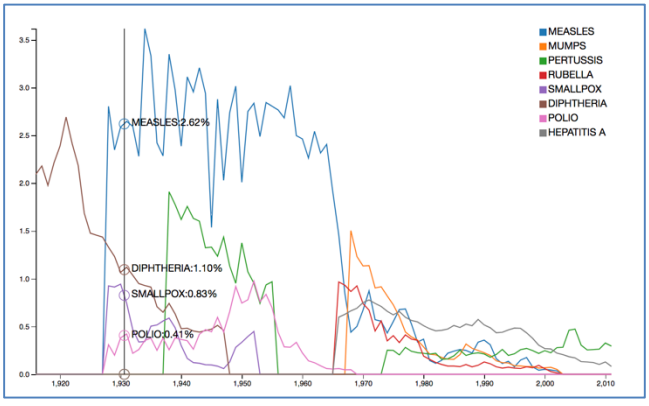

Figure 2: *Multiline chart, trends of diseases*

Layout: multiline chart.
Visual encoding: x-axis represents the time (aggregate the original data by year); y-axis represents the incidence; color of each line represents the category of disease.
Interaction: mouse over will show the details of each diseases in a year.
Discussion: as we see in Figure 2, all the diseases are in a decrease tendency in these 100 years. In 1940s and 1970s, these are two big disease explosions.

Other than that, since the contagious diseases' preventions could also be influenced by the economic development significantly, we also plan to design a visualization to show the relationship between diseases and economic development, where the economic development could be represented by GDP per capita. Scatterplot is one of the best way to visualize relationship. We design this visualization as follow:

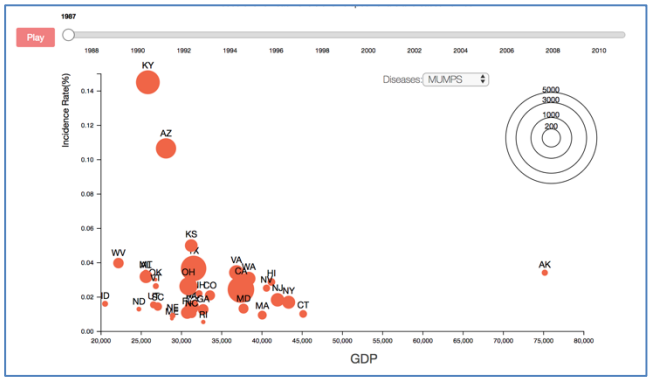

Figure 3: *Scatterplot, relationship between diseases and economy*

Layout: scatterplot.
Visual encoding: x-axis represents GDP per capital; y-axis represents the incidence; size of each scatter represents cases; text labels show the state names.
Interaction: drop down box to select among diseases; slide to select among years; and there is also a button to play the changes through the whole time interval.
Discussion: as we expected, the scatters move from top-left to bottom-right over time, which suggests that diseases have a negative relationship with economic development, though there are several explosions. Low GDP states are more vulnerable to contagious diseases.

4.2 Seasonal Patterns

It is well known that some of epidemic diseases have significant seasonal cycles. That why we need to get influenza vaccines every fall. Whether this kind of seasonal cycles exist and when the peaks of a year are going to happen are very important problems in contagious diseases' preventions: if we know the answers to these problems, we will be able to deal with these diseases more accurately and effectively, just like what we are doing to influenza. Long term historical data is a good source to reveal this kind of cycles.

In data visualization techniques, heap map is the most popular way to show clusters or peaks. Since we are dealing with cyclic time oriented data, we will use a circular heat map to find the seasonal patterns in diseases. Compare to traditional matrix heat map, circular heat map could reveal circular patterns more easily even the peaks appear at the beginning or ending of each time round. We design this visualization as follow:

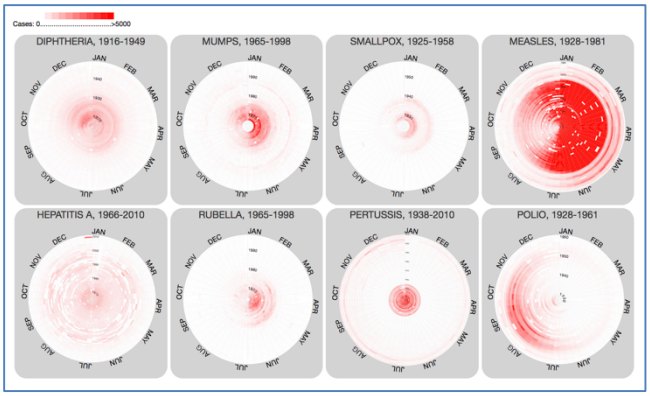

Figure 4: *Multiple circular heat maps, seasonal patterns of diseases*

Layout: multiple circular maps.
Visual encoding: the angle represents the same week of each year; the radius represents the year; the color of each grid represents cases of disease in the whole country per week.
Interaction: click each small circle will give a single large circular heat map of the disease. Mouse over on the large figure will show the details of each grid.
Discussion: as we see in Figure 4, we can not only study on the seasonal pattern of each disease but also compare across different diseases. We can find that there is no significant seasonal pattern in HEPATITIS A and PERTUSSIS; the peak of DIPHTHERIA appeared in Nov – Dec; most cases of MUMPS and SMALLPOX appears in the first half of a year; the peak of RUBELLA appears in Apr – Jul; the peak of POLIO appears in Aug – Oct; and there will be less of MEASLES during Aug – Nov. Due to the data availability issue, each circular heat map covers different time interval. Another problem of this visualization is that each states could have different seasonal patterns due to local environments, but since we sum cases from all states together, we cannot show the difference across state.

4.3 Spatial Patterns

Respecting the geographic locations, some places are more vulnerable to some diseases. For example, since Zika is spread by mosquitoes, it is more dangerous in Florida than in Pennsylvania. This pattern is also clear in historical data of contagious diseases.

We design a geographical map to show this pattern. Using the map we will be able to connect disease data and its spatial information, compare the incidences across different stats and explore the spatial influence. The map shown as follow:

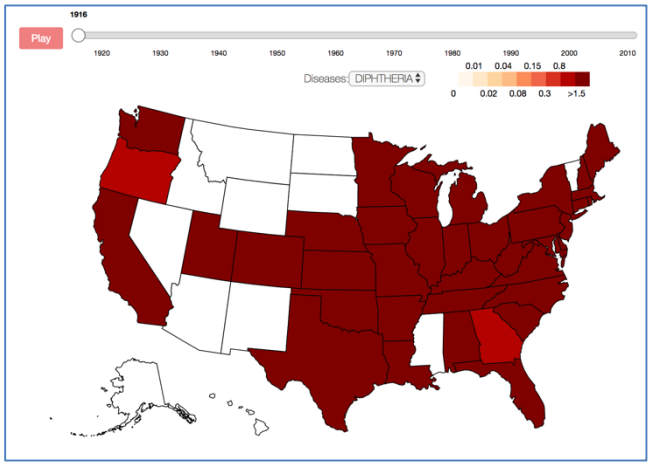

Figure 5: *Map, spatial patterns of diseases*

Layout: geographical map.
Visual encoding: 2D position represents the geo-spatial location of each state; the color of each state represents the incidence.
Interaction: drop down box to select among diseases; slide to select among years; and there is also a button to play the changes through the whole time interval. Mouse over the map will show the detail information of each state in a year, and mouse over on legend will highlight states of that value.
Discussion: we found from this visualization that spatial locations do have some influence on contagious diseases. For example, DIPHTHERIA has a higher incidence at both east and west coast but lower rate in central region. This pattern could be caused by the high population concentration in east and west coast or other potential factors like weather.

The map is a good way to overview the spatial patterns, but we want to go further to see and compare the trends in every state. Therefore, we plan to design a small multiple as follow:

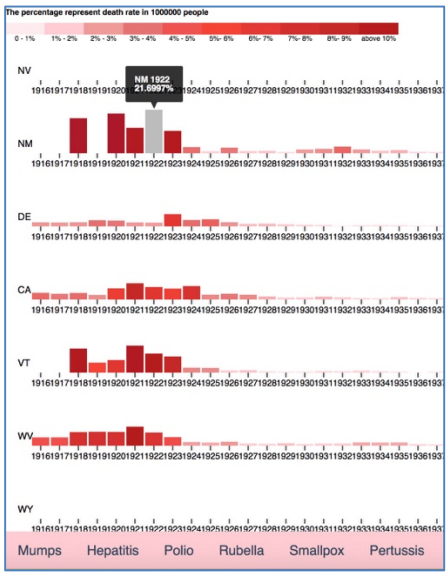

Figure 6: *Small multiple bar charts, comparing trends across states*

Layout: small multiple bar charts.
Visual encoding: 2D position represents the geo-spatial location of each state; the color of each state represents the incidence. x-axis represents the time (aggregate the original data by year); y-axis and the color of each bar both represent the incidence.
Interaction: menu bar to choose among diseases. Mouse over show the detail information of each bar.
Discussion: this small multiple can show the trends in every state as well as comparison across states. With this visualization, we will be able to identify the spatial location of each explosions. We have no enough time to include this visualization in our demo, but we believe that this figure could help to show the trends and patterns of contagious diseases.

5 EVALUATION, DISCUSSION AND FUTURE WORK

We assume that a typical user of our project could be people without academic or professional background in medical science and contagious diseases, but want to know more about this topic. We hope our designs could tell the user about basic trends and patterns in contagious diseases' spreads and preventions.

A simple interview which has only one respondent suggests that the user was able to understand the trends and patterns we want to show through the visualizations easily, but it was hard to get more information about why these trends and patterns happened like this, which is also interested by the user. To improve that, we may need to add more annotations and explanations. Also, we plan to get more information about a brief history of these contagious diseases, and use it as narrative to show a timeline and do a better story telling.

Another issue reported by the user is that we did not organize all the visualizations well to convey our ideas. Actually, we planned to link the visualizations in the same topic but we did not have enough time to implement it. As we planned, we will use a same time slide to control the time selection in multiline chart and scatterplot, so that the user will be able to understand the relationship of economy and diseases in this year, as well as the position of this year in the whole trend. Similarly, we planned to link the geographical map and the small multiple bar charts, so that when the user select and highlight a states in the map, the bar chart of the state will also be highlighted. Through this linkage, the user will be able to see the trend of the state as well as the position of the state in the whole country in a year.

Besides the improvements we mentioned above, we may also apply the same designs using Project Tycho level 2 data, which contains the information of more diseases. Through this work we may be able to show a clearer picture of contagious diseases.

6 CONCLUSIONS

In this project, we aim to reveal the general trends, seasonal patterns and spatial patterns of contagious diseases in the US using visualization designs. Through our designs, we found that all the diseases are in a decrease tendency in the past 100 years and have a negative relationship with economic development, though there are several explosions. There are at least 6 diseases have clear seasonal patterns. And spatial locations also have significant influence to the diseases. All these trends and patterns appear as our expected from our assumptions. On the other hand,

pivot user study shows that our designs can show the trends and patterns clearly to the user. With some improvements, our designs will be able to tell a good story of the history of the contagious diseases in the US.

ACKNOWLEDGMENTS

Jingran preprocessed the data, completed the multiline chart, the scatterplot and the map. Chen finished the circular heat maps and drafted the report. Jiexiao did the small multiples for spatial pattern.

REFERENCES

[1] Yasuko Matsubara, Yasushi Sakurai, Willem G. van Panhuis, and Christos Faloutsos. 2014. FUNNEL: automatic mining of spatially coevolving epidemics. *In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '14)*. ACM, New York, NY, USA, 105-114. DOI: https://doi.org/10.1145/2623330.2623624.

[2] Van Panhuis WG, Grefenstette J, Jung SY, et al. Contagious Diseases in the United States from 1888 to the Present. *The New England journal of medicine*. 2013;369(22):2152-2158. doi:10.1056/NEJMms1215400.

[3] Project Tycho: https://www.tycho.pitt.edu/.

[4] GDP per capita: Bureau of Economic Analysis (https://bea.gov).

[5] Wikipedia: Black Death https://en.wikipedia.org/wiki/Black_Death.

[6] Wikipedia: Smallpox vaccine, https://en.wikipedia.org/wiki/Smallpox_vaccine.