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

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

1 INTRODUCTION

In the history of medicine and also 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.

According to Y. Matsubara et al. (2014) and W. G. Panhuis et al. (2013), there could be at least three kinds of basic trends and patterns in the history of contagious diseases spreads and preventions: (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.

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 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. (About Project Tycho Data: https://www.tycho.pitt.edu/about.php) 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. (Level 1 data: https://www.tycho.pitt.edu/data/level1.php) 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 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.

3 VISUALIZATION DESIGN

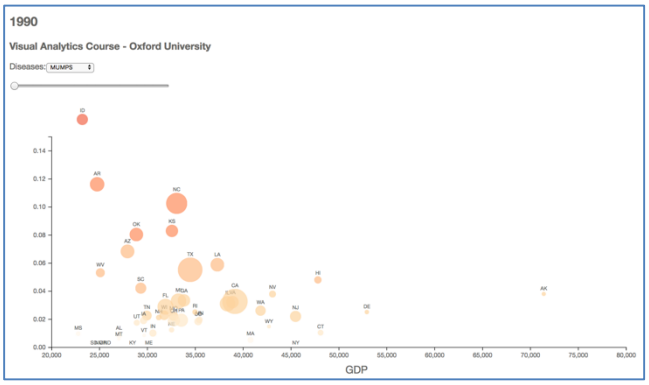
3.1 Trends in the past 100 years

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, https://en.wikipedia.org/wiki/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.

To visualize a time-series data, line chart or index chart is definitely our first choice, depending on which kind of data we are visualizing (count or rate). This chart is still in proceeding.

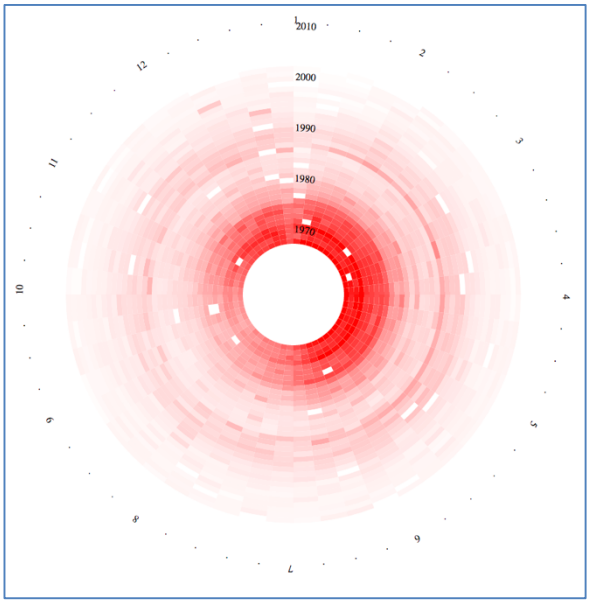
Other then that, since the contagious diseases' preventions could also be influenced by the economic development significantly, we also plan to design a visualization to show the relationship between diseases and economic development, where the economic development could be represented by GDP per capita. We choose scatterplot to show the relationship, and the prototype of the visualization is shown as follow:



3.2 Seasonal cycles

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.

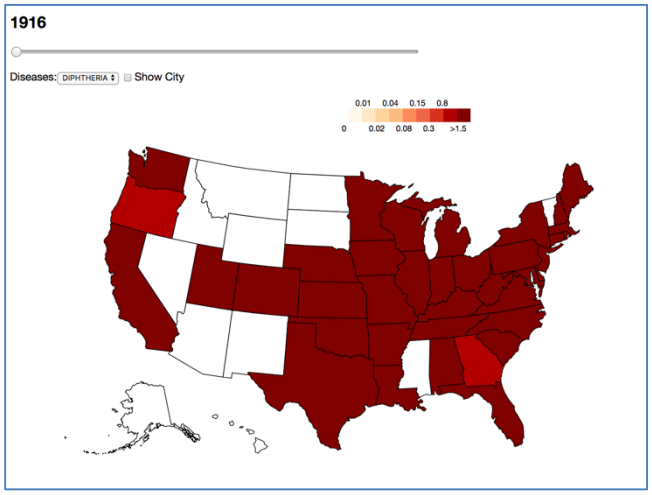
In data visualization techniques, circular or radial charts are most frequently used to discover these cycles in cyclic time oriented data from, i.e. circular silhouette, circular heat map and radial line chart. In this project, we first tried circular heat map which is shown as below. We can see from this chart that the spread of Diphtheria is slower during summer (July to October).



3.3 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.

To show this pattern, we design a map which uses different color to represent the incidences of each state in a specific year, and uses a slide to change from time to time. The prototype of the visualization is shown as follow:



4 EVALUATION AND DISCUSSION

In the next two weeks, we will finish the charts in proceeding mentioned above, annotate existed chart more clearly, and integrate all charts into a whole picture. We also plan to include a pilot user study to evaluate our designs.

5 CONCLUSION

In proceeding.

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.