Journal of World-Historical Information
JWHI

REVIEWS OF DATASETS

Patrick Manning

Large-scale data on U.S. disease, 1887 - 2014

Review of: Project Tycho: Data for Health

Author/Compiler: Wilbert van Panhuis et al.

Repository: http://www.tycho.pitt.edu

Funded by: The Bill & Melinda Gates Foundation and the National Institutes of Health

Date of Posting: September 2013

Size: unknown

Volume 2-3, No. 1 (2014-2015) | ISSN 2169-0812 (online)
DOI 10.5195/jwhi.2015.24 | http://jwhi.pitt.edu

The Tycho Project displays systematic data on disease surveillance throughout the United States from 1887 to 2014. It is named after the sixteenth-century Danish astronomer Tycho Brahe, in honor of his systematic collection of data on the positions of heavenly bodies. Work was completed at the Graduate School of Public Health and the Center for Global Health at the University of Pittsburgh. Initial project design was inspired by Donald Burke; the construction of the dataset from raw data was led by Wilbert van Panhuis.

The dataset includes over 14 million observations on cases or deaths from any of 58 diseases recorded in localities throughout the United States starting 1888. The "surveillance" procedure consisted of keeping track of numbers of cases or deaths from identifiable infectious diseases and submitting them to local and national health authorities. The Tycho website is a comprehensive introduction to this large-scale dataset. Tycho provides access to the data itself, ways to use the data, and a full range of background materials. The "Home" and "About" pages provide an overview of the project, to be discussed below. The "Data" page gives access to three levels of detailed information on levels of disease incidence by year and by city or state. Selection of any combination retrieves data for these criteria in the form of a graph, a table, and an Excel sheet. The "Resources" page lists citations, images, tools for analysis. Additional links go to Frequently Asked Questions, News, and Contacts.

In an elegant diagram providing an overview of the data, a "circle graph" presents the whole period from 1888 to 2010, displaying all 58 of the diseases recorded, and showing which of them were documented for which year. This one image provides a whole history of health and medical treatment in the United States. Thus, typhoid fever was recorded from 1888, mumps began to be recorded in roughly 1920, and AIDS began to be recorded after 1980; tetanus was recorded only up to 1980.

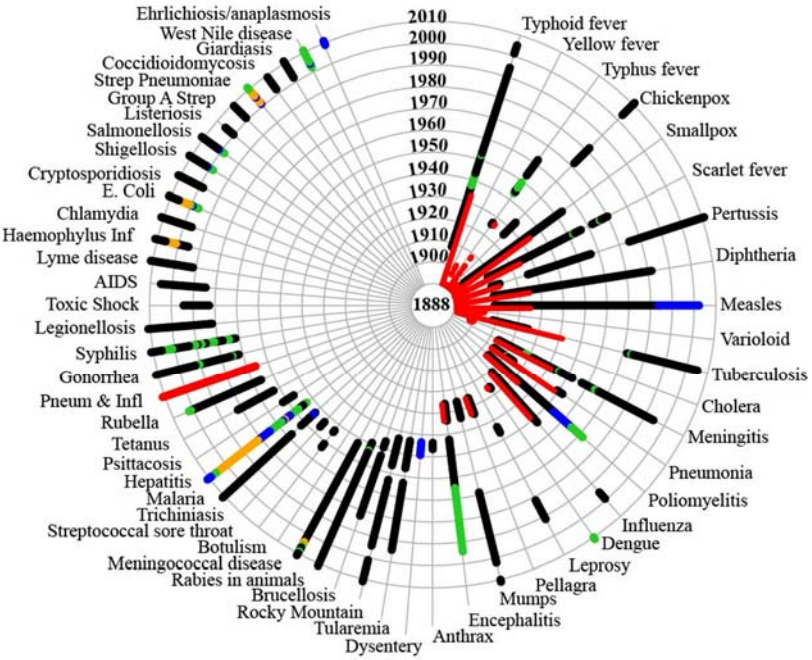

Figure 1. "Circle Graph"

The data in Tycho result from completed digitization of the entire history of weekly National Notifiable Disease Surveillance System (NNDSS) reports for the United States (1888-2013) into a database in computable format. The team at the University of Pittsburgh collected, digitized, and processed the data, in work that involved the challenging tasks of locating the dispersed original print volumes and double-entry plus confirmation of all the data (conducted by university students in Southeast Asia). Following digitization, which took place under the inspiration and leadership of Donald Burke, Wilbert van Panhuis led the process of cleaning, documenting, and organizing the data. The data are now organized at three levels of documentation and consistency. Overall the results address as many as 58 diseases; 8 of these have been selected for close analysis at the first stage; further documentation and analysis continues.

Exploring the data

Level 1, the most fully documented data, includes materials standardized in a common format. Level 1 includes data on eight diseases and includes 0.7 million "counts" or reports from 122 cities. – different types of counts, standardized in a common format (0.7 million, 8 diseases, 122 cities) These data are best placed for detailed analysis. Level 2 includes a much larger number of data for which a common format has been achieved but which are not yet standardized. This includes 3.6 million counts of data on 50 diseases from 1284 cities. These data are also open for analysis through the Tycho interface. Level 3 is that of the raw data after its basic processing. It includes 14 million counts on 58 diseases, from 3026 cities.

The "data count" is a variable designed by the Tycho staff for reporting health data. It is a key example of the original accounting work that was necessary to build this data resource. A "count" is defined as the number of cases or deaths due to a disease in a specific location and time period. A count is equivalent to a data point in the dataset. One count is not a single infected individual, but a single observation on the number of infected individuals in a given time and place. As the "About" page indicates, as many as 17 categories of "count" have been identified, ranging from counts from counties, from cities, counts for data not yet harmonized, and counts containing non-disease information.

To explore the data, it makes sense for users to start with Level 1 on the "Data" page. There the user may easily see an overview of the total counts for the eight diseases. The diseases are measles, pertussis, hepatitis A, polio, mumps, rubella, smallpox, and diphtheria. The user may use an interface to explore the data or to recall data for one disease at a time. For instance, selecting hepatitis A for Kentucky for the years 1966-2013 instantly brings up a page of results. The results begin with a graph showing weekly cases of hepatitis A for the whole period, accompanied by a chart showing the same data. An adjoining link makes it possible to download an Excel file containing the data. Excel files from multiple searches could be combined to permit larger scale analysis.

The interface for Level 2 displays 50 diseases. Data for them may be retrieved in a fashion parallel to Level 1. In addition, a link to an API makes it possible to download Level 2 data in this fashion.

For Level 3, the data cannot be downloaded so neatly. Nevertheless, researchers can request raw data from Level 3 by region, time period, and disease. In addition, for researchers who want to do large-scale work, it is possible to obtain a data dump.

Representing the data

Among the strengths of the Tycho Project are excellent representations of the data. The "circle graph," displayed above, gives a broad and chronological overview of the surveillance data. Second, chronological graphs provide peaks showing the weakly incidence of disease for each disease, region, and time frame. These graphs are helpful in revealing the seasonal characteristics of some diseases (with annual peaks) as well as the long-term rise and fall. Of particular interest are the changes in disease incidence in times following the implementation of a new vaccine.

A third type of display is the "heat map." This graphic device lists time along the x-axis and lists the many places that are reported along the y-axis. (Usually the reports are by state.) The data for each week and each state are represented by color, with deep blue for the highest counts and light yellow for the lowest counts. The result gives a detailed national picture, showing the fluctuations and variations in incidence by time and place. These heat maps are especially useful in showing the impact of new vaccines. Thus the measles vaccine brought a rapid and nearly complete decline in measles incidence within two or three years; the pertussis vaccine (which was not as effective and especially not as popular) brought a decline in outbreaks, but substantial counts of pertussis continued for years after the vaccine appeared.

Resources

The Tycho Project website includes a rather thorough set of supported resources. It includes links to the projects major academic papers, especially its initial paper in the New England Journal of Medicine: Van Panhuis, et. al., Contagious Diseases in the United States from 1888 to the present. *NEJM* 2013; 369(22): 2152-2158. Other resources, intended to spread use of the website, include downloadable images (including a full PowerPoint), a brochure suitable for print display, and videos. The video tutorial walks users through all steps of the site; the animation of data conveys – especially the tutorial, animation of data, R-language code for generating heat maps and circle diagrams. Also included is an extensive set of preliminary reports on historical disease patterns for U.S. states.

Data Collection and Methods

The website emphasizes display and usability of data. As such, it rather downplays the considerable labor that was involved in collecting and documenting the data. The original data consist of weekly reports by health officials on the number of cases or the number of diseases in cities, counties, and states throughout the United States. These data, assembled and published by successive organizations of the U.S. government, were held irregularly in libraries. The initial task was locating all of the originally published data. The next step was digitizing the data (a process conducted through double-entry by university students in Cambodia and Laos).

In sum, the Tycho Project is a remarkable example of a successful data project. It was initiated with the vision of "data rescue" and print documents were scheduled for shredding. It was built by a persistent and skilled team. The project gained foundation support, a prominent placement in its initial roll-out, and an ongoing link to the *Wall Street Journal.* There is still more to do. Indeed, the Home page of the site gives attention to advocacy activities and to projects in development, on dengue and chikungunya. In addition, the project is thorough in giving recognition to the full range of staff and researchers who have carried out work on it.

This journal is published by the University Library System of the University of Pittsburgh as part of its D-Scribe Digital Publishing Program and is cosponsored by the University of Pittsburgh Press.