Mid Term Project Steps and Challenges

Carla Harrell

Bellevue University

The topic detail has been selected on personal interest and the data has been collected from online resources mentioned in the reference section. It was chosen because I work in clinical laboratory and have always been curious on the mandates reporting disease outbreaks. I just selected one of the tables from the disease outbreaks which consisted of gonorrhea, Haemophilus Influenza Invasive Disease. The law requires monitoring of the disease to prevent outbreaks. The dataset contains provisional cases of national notifiable diseases from the National Notifiable Diseases Surveillance System (NNDSS). NNDSS data from the 50 states, New York City, the District of Columbia and the U.S. territories are collated and published weekly on the NNDSS Data and Statistics web page (<https://wwwn.cdc.gov/nndss/data-and-statistics.html>). Cases reported by state health departments to CDC for weekly publication are provisional because of the time needed to complete case follow-up. Therefore, numbers presented in later weeks may reflect changes made to these counts as additional information becomes available. The national surveillance case definitions used to define a case are available on the NNDSS web site at <https://wwwn.cdc.gov/nndss/>. Information about the weekly provisional data and guides to interpreting data are available at: <https://wwwn.cdc.gov/nndss/infectious-tables.html>. Data are finalized 10 months after the end of the year.

**U: Unavailable** — The reporting jurisdiction was unable to send the data to CDC or CDC was unable to process the data.

**-**: **No reported cases** — The reporting jurisdiction did not submit any cases to CDC.

**N**: **Not reportable** — The disease or condition was not reportable by law, statute, or regulation in the reporting jurisdiction.

**NN**: **Not nationally notifiable** — This condition was not designated as being nationally notifiable.

**NP**: **Nationally notifiable but not published** — CDC does not have data because of changes in how conditions are categorized.

**Cum**: **Cumulative** year-to-date counts.

**Max**: **Maximum** — Maximum case count during the previous 52 weeks.

**\*** Case counts for reporting years 2018 and 2019 are provisional and subject to change. Cases are assigned to the reporting jurisdiction submitting the case to NNDSS, if the case's country of usual residence is the US, a US territory, unknown, or null (i.e. country not reported); otherwise, the case is assigned to the 'Non-US Residents' category. For further information on interpretation of these data, see

**Notes on Data:**

**Reporting area** – This column represents the U.S. Department of Health and Human Services Regions (HHS) and the jurisdictions (50 U.S. states, five U.S. territories, New York City and Washington DC) that submit case notifications to NNDSS. The non-US residents’ data represents illness that occurred in the US in individuals whose country of usual residence is outside the US or US territories.

**Week** – The week identified in this table refers to the variable “MMWR week” which represents the week of the epidemiologic year (MMWR year) for which the NNDSS disease report is assigned by the local, county, or state health department, for the purposes of disease incidence reporting or publishing..

**Current Week** – For a case to be published in the table under current week, it must have been reported to CDC during that week and assigned by the jurisdiction to that MMWR week.

**Previous 52 weeks Max**– To calculate the national maximum, the data are first aggregated to the national level for each week, and the maximum number of cases for the year-to-date is identified.

**Variables:**

* **MMWR Week**: Represents the week of the epidemiologic year for which disease report is assigned by local, county, or state health department for purposes of disease reporting.
* **Gonorrhea Current Week flag**:
* **Gonorrhea Cumulative for 2019**: The year to date provisional counts for Gonorrhea in 2019
* **Gonorrhea Cumulative for 2019 flag**:
* **Gonorrhea Cumulative for 2018:** The year to date provisional counts for Gonorrhea in 2018
* **Gonorrhea Cumulative for 2018 flag:**
* **Haemophilus Influenzae\_Invasive Disease\_All Ages\_All Serotypes Current Week:** The number of cases reported weekly
* **Haemophilus Influenzae\_Invasive Disease\_All Ages\_All Serotypes Current Week Flag:**
* **Haemophilus Influenzae\_Invasive Disease\_All Ages\_All Serotypes Previous 52 Weeks Max:** The Maximum of cases reported at 52 weeks
* **Haemophilus Influenzae\_Invasive Disease\_All Ages\_All Serotypes Previous 52 Weeks Max flag:**
* **Haemophilus Influenzae\_Invasive Disease\_All Ages\_All Serotypes Previous 52 Weeks Maximum**: The Maximum of cases reported at 52 weeks
* **Haemophilus Influenzae\_Invasive Disease\_All Ages\_All Serotypes Current Week flag:**
* **Haemophilus Influenzae\_Invasive Disease\_All Ages\_All Serotypes Cumulative For 2019:** The year to date provisional counts for Gonorrhea in 2019
* **Haemophilus Influenzae\_Invasive Disease\_All Ages\_All Serotypes Cumulative For 2019 flag:**
* **Haemophilus Influenzae\_Invasive Disease\_All Ages\_All Serotypes Cumulative For 2018:**  The year to date provisional counts for Gonorrhea in 2018
* **Haemophilus Influenzae\_Invasive Disease\_All Ages\_All Serotypes Cumulative For 2018 flag:**

**Mid Term Project Steps and Challenges**

The data for the midterm project is collected from the CDC. The data file location is: https://data.cdc.gov/NNDSS/NNDSS-TABLE-1M-Gonorrhea-to-Haemophilus-influenzae/h4wb-nae4 and the file names is as follows:

“NNDSS\_\_TABLE\_1M.\_Gonorrhea\_to\_Haemophilus\_influenzae\_invasive\_disease\_age\_5\_years\_Serotype\_b. “

**Step 1: Import Data**

The data files are read into ‘pandas’ dataframe using pandas ‘read\_csv’ method.

**Step 2: Replace Header**

The headers were already having proper understandable format to understand. In order to tweak them, the headers there were some symbols and other punctuations in them after importing which were removed

**Step 3: Outliers**

I was able to find from the five-point summary of the variables by using describe function of pandas. Then, a 98th percentile was used to seal the numeric variables to treat the outliers.

**Step 4: Finding and Handling Duplicate Rows**

Since there were no duplicates in the dataset, there was not so much we could do in this section.

.

**Step 5: Fuzzy Matching**

Challenge lied in identifying the variable to conduct fuzzy matching. There are no two 'logically making sense variables' in the dataset on which fuzzy matching could be done. However, I then considered to do a cross join of location variable in the dataset and find out the fuzzy scores with other locations in the same variable. Then, it was a straightforward thing to have it printed for fuzzy scores>50.

I am still trying to understand the variables that has the word flag included and will email the dataset owner.

**References:**

https://data.cdc.gov/NNDSS/NNDSS-TABLE-1M-Gonorrhea-to-Haemophilus-influenzae/h4wb-nae4

Readers’ Guide: Understanding Weekly and Annual National Notifiable Diseases Surveillance System WONDER Tables. https://wwwn.cdc.gov/nndss/document/Users\_guide\_WONDER\_tables\_cleared\_final.pdf