**DATA624 - DATATHON #4**

**GROUP 3 - HANNAH DIMMICK, MINKI LEE, ADAEZE NWIGWE**

**TRACKING DISEASE OUTBREAKS IN THE UNITED STATES, 1971-2018**

**Introduction:**

Having a centralized method of accounting for disease outbreaks is critical information for public health departments in order to protect the public. The United States Center for Disease Control (CDC) maintains the National Outbreak Reporting System (NORS) dashboard in order to "enhance national outbreak reporting [1].” This is imperative to improving surveillance and creating policy to avoid future outbreaks and trace possible affected individuals. The data is collected via reports from local, state, and national public health agencies and integrated into the publicly accessible NORS dashboard in order to help provide high quality information about these outbreaks.

The dashboard includes reports of 2+ cases of enteric and non-enteric illnesses as the result of a common exposure. These exposures can be from bacteria, viruses, parasites, chemicals, or toxins delivered by food, water, animals, other people, or environmental contaminations. NORS also tracks outbreaks caused by unknown modes of transmission/causes of illness that appear to have a common origin and cause illness in 2+ individuals.

NORS has collected data on waterborne outbreaks since 1971, foodborne outbreaks since 1998, and other outbreaks since 2009. The centralized nature of this dashboard is incredibly important in identifying trends in disease and opportunities to improve public health measures. Evidence of the need for these types of systems has been recently demonstrated in the COVID-19 pandemic, where the ability to quickly track and centralize information about outbreak incidences has been extremely valuable and time sensitive [2].

To further investigate the nature of outbreaks in the United States, we used data science and visualization techniques to develop meaningful information from the NORS dashboard data. Our guiding questions for the project were:

*Question 1: Which state had the most infections during a given time period? What types of transmission modes were most common in each state? And how deadly were each of these transmission modes?*

Different geography, agricultural industries, and population densities between states presumably lead to different rates of outbreaks of certain kinds. Additionally, certain types of modes are more likely to produce serious illness and death. Our team was interested in getting an overview of these differences.

*Question 2: Which year has the highest number of deaths, and what is causing death?*

*Question 3: In the year with the highest number of deaths, in what setting are people usually exposed the top 5 leading causes of death?*

Certain type of infections may be more easily acquired in certain settings. Our team wanted to better understand where individuals are exposed to infections and in what settings do different infections spread, to see if there are any observable patterns.  
**Dataset:**

The dataset used was “Food and waterborne infections” from the CDC NORS dashboard. The dataset lists the month/year of each infection, location, main mode of transmission, and details about strains, illness rates, death rates, and whether an illness was confirmed or suspected.

**Analysis:**

*Question 1*: An exploratory visualization was created to gain an overview of information on number of illnesses, deaths, and mortality rate of certain transmission modes (percent of illnesses resulting in death). This involved assigning each state’s information to a map, which functioned as a filter to create a bar graph of illness and death counts for each transmission mode. Each mode was colored according to the mortality rate in order to quickly gauge how deadly a particular transmission mode was. The year range was set to 2009-2017, as this is the only time period for which information on all transmission modes were collected. Rates of highest illness, death, and mortality rate were calculated via Microsoft Excel.

*Question 2:* A line chart was created to show the yearly distribution of deaths by all causes and find which year has the highest number of deaths. Any null values were excluded from the visualization. After finding the year with the highest number of deaths, outbreak data were filtered using Python to investigate the top 5 leading causes of deaths. Unnecessary columns, such as "Setting" and "Month", were removed, and only the columns of interest, such as "Year" and "Deaths", were extracted. Missing information on the "Etiology" column is dropped, and any null values in the "Deaths" column were replaced with zeros. We counted the total number of deaths for each of the top 5 leading causes and put it in a new data frame.

*Question 3:* A bubble chart was created to show setting of exposure for the top 5 leading causes of death in 2014. Clean up and preparation of the dataset to answer this question was done through R. The setting column from the dataset was categorized and cleaned up to narrow down the types of settings of exposure; for example, all restaurant types such as fast-food or sit down-dining, where categorized under restaurant. The dataset was filtered to only include the year 2014 and only include the top 5 leading causes of death. Columns that were not necessary for answering the question and rows with missing etiology, setting, or illness data were dropped. The data was grouped by Etiology and Setting, and a calculated percentage field was added to the data frame. Etiology, setting and percentage of illnesses are presented in the bubble chart.

**Findings:**

*Question 1:* The state reporting the highest number of deaths for the given time period was Wisconsin, reporting 282 deaths across 63,435 illnesses. The main source of the deaths occurred via person-to-person contact, which was responsible for 192 deaths (out of 56,074 illnesses). The state with the most reported illnesses for the time period was New York, with 74,011 reports. This mainly came from person-to-person contact (68,292 illnesses). However, the highest overall mortality rate for the time period was the state of Mississippi, which reported 0.411% of illnesses resulting in death in the given time period, with 16 deaths due to person-to-person contact (3,365 illnesses) and 1 death due to waterborne (9 illnesses).

In the entire United States, the greatest transmission mode was person-to-person contact (responsible for 685,341 illnesses in the given time period), which was also responsible for the most deaths (842). However, the transmission mode associated with the highest mortality rate was water, which had 0.928% of illnesses resulting in death.

*Question 2:* When we look at the line chart, we see a rapid increase in deaths between 2009 to 2011. According to Kathy Benedict, an epidemiologist in CDC, the rapid increase in reported deaths may be due to an actual increase in disease, the development of NORS in 2009 which allowed more information to be collected on the number of deaths from waterborne and foodborne disease outbreaks, and changes in surveillance capacity in states. The total number of reported deaths from 1971 to 2018 is 1805; the top 5 leading causes of death were: Norovirus, Legionella, Listeria, Salmonella, and Escherichia. We observed the highest number of deaths in 2014, and norovirus was the most common cause of the outbreak.

*Question 3:* The bubble chart shows that setting of exposure varies from one etiology to another. Some of the most common settings for exposure are care facilities and private residence. Restaurants is a common place of exposure for a bacterium such as salmonella, but not a common setting of exposure for other food and water bore infections.

**Conclusions:**

*Question 1*: The extremely varied results in illnesses, deaths, and mortality rates between states seem to imply that a variety of local factors are at play in determining what causes outbreaks of illness and death. It is difficult to draw concrete conclusions about the factors that may have caused particular states to have higher levels of particular pathogens or higher levels of disease overall. An interesting outlier in the data was Wisconsin, which seemed to have a disproportionate number of deaths. It is only the 20th most populous state but accounted for an outsized number of severe illness. A closer look at the data reveals several particularly bad Norovirus outbreaks in long-term care facilities/nursing homes in the 2010s that are responsible for a significant number of these reported person-to-person illnesses and deaths.

Interestingly, water was the outbreak mode most likely to result in death in the whole country. A closer look reveals that this is largely due to a few states that report high numbers of waterborne-related deaths (Illinois, Ohio, Pennsylvania, New York, and Michigan), which have significant water borders. However, since waterborne outbreaks can be reported in both non-naturally occurring bodies of water (i.e., swimming pools, fountains) and land-locked bodies of water (i.e., lakes, ponds) in addition to coastal bodies of water, more analysis is required to determine if the association with waterborne-related deaths is due to illnesses contracted in coastal bodies of water or if it is coincidental.

*Question 2*: Although deaths are rare from the diseases, we observed many deaths despite the rapid evolution of the scientific and technological advances in infectious diseases research. Health and food industry professionals can use our findings to target innovative prevention efforts related to pathogens and foods that cause waterborne and foodborne disease outbreaks. Furthermore, to maintain and improve public health, they can connect with community groups to promote disease prevention and awareness programs.

*Question 3: V*arious food and waterborne infections spread in different settings. While we expected to see high exposure related to restaurants or caterers or banquet facilities, these are usual places of exposure just for few of the infections causing death. Salmonella for example is one of the infections that had high exposure in restaurant settings. This can be explained by considering the nature of the bacterium and the way it is transmitted, which is usually through eating uncooked animal products. Aside from Salmonella, exposure to infections in restaurants is low. This is probably because restaurant facilities, and businesses in the food industry must respect stringent protocols and often undergo inspections to ensure that food is prepared and stored appropriately. Care facilities and private residences on the other hand, which were common setting for exposure to infections, are not subject to the same regulations as business in the food related industries hence the higher number of exposures to infections in these settings.

**Limitations:**

The NORS dashboard does not include all data reported through NORS; the full dataset can only be accessed by specific request, therefore the information used in this analysis is incomplete even before taking into account potential issues with/lack of reporting from certain jurisdictions [1].

Conclusions about root causes of differences between illness/death/mortality rates in different states were essentially impossible to draw. Additionally, clusters were hard to identify as no cities were provided (i.e., it would have been interesting to determine if the yearly Norovirus outbreaks in Wisconsin were all in the same location or if they were spread out across the state – either answer would provide valuable information). Further analysis should be done to look into possible community clusters. Additionally, the mortality rate measure is more valuable when looking at large sample sizes than smaller ones. For example, seeing that food pathogens in California have a 0.138% mortality rate (14 deaths per 10,173 illnesses) is much more valid information about the nature of foodborne illness than the fact that environmental contagions in Kansas have a 9.09% mortality rate (1 death in 11 illnesses).

Setting of exposure information was missing from a considerable portion of the data which limits the reliability of the analysis. In addition, the setting field is self-reported which means it is likely not a very accurate or reliable source. It may be hard for an individual to correctly identify how they acquired an infection as an individual may not remember or may be unsure of exactly what caused the illness.

**References:**

[1] National outbreak reporting SYSTEM (NORS). (2018, December 7). Retrieved February 28, 2021, from <https://wwwn.cdc.gov/norsdashboard/>

[2] Feng, C., Lee, T., & Tam, D. (2020, November). *Disease Outbreak Radar: A Tool for Epidemiologists* [Scholarly project]. In *cs.ubc.ca*. Retrieved February 28, 2021, from <https://www.cs.ubc.ca/~tmm/courses/547-20/projects/cloris-derek-harry/proposal.pdf>