Exercise 1.4 Task – Sourcing the Right Data

Hypothesis: "*If a citizen is 65 years of age or older, then they are more at risk of developing serious complications, including death, from the flu.*"

1. US Census Bureau – Population Data by Geography, Time, Age, and gender

Data Sourcing:

This is external data sourced from the US Census Bureau. This is government data used to assist in the planning of various government sectors and services across the country. As such, it is held to high standards of authenticity and should be considered a trustworthy data source.

Data Collection:

This data is administrative in nature. The data within this dataset is sourced from other agencies' electronic records such as birth certificates and combined with census and survey data. Hence, the collection method is primarily manual. Time lags are also very likely to have affected this data depending on the frequency of data gathering and verification by the US Census Bureau and the various agencies that data may have been sourced from.

Data Contents:

The dataset contains yearly population counts for each county in each US continent between the years 2009 and 2017. Population counts for each county are divided into various categories such as *5-year age groups* and *gender*.

Data Limitations:

The most obvious limitation to this data is that it only covers 2009 to 2017. This means that any demographic changes since 2017 will not be accounted for. This is a significant gap that will likely affect any future projections.

Manual entry errors are also a possible limitation of this data set. Although a high level of authenticity and rigor can be assumed due to the source, multiple collection methods and need for human typing present a likelihood of various errors.

Data Relevance:

This data set provides highly relevant demographic information for the project. The total population counts in all regions will indicate where more medical professionals will be required across the US. Moreover, the data permits specific insights to be made regarding the distribution of vulnerable populations, specifically, individuals over 65 years of age. This is highly relevant to the project as it is assumed medical professionals will be more urgently required in states with larger vulnerable populations.

On its own, this data set will not be enough to test the hypothesis about higher mortality rates for individuals over 65 years of age. However, it can be analyzed alongside data about hospital admissions and influenza mortality rates by state and age to determine if this vulnerable population does experience higher mortality rates. By comparing relative mortality rates and/or hospital admissions of elderly people in each state, a staffing plan can be more confidently constructed for the project. For these reasons, this data set is critical to the project objective.

2. CDC – Influenza Visits and Lab Tests Data Sets

Data Sourcing:

This is external data compiled by and made publicly available by the Centers for Disease Control and Prevention (CDC). The CDC gathers and maintains this data from many outpatient health care providers, public health providers, and clinical laboratories. Both data sets should be considered trustworthy because the CDC is a reputable government organization, and the data is reported by an expansive network of health professionals.

Data Collection:

Both data sets are examples of survey data because they do not display the total count of all patient visits with ILI or lab tests conducted across the country. This is to be expected due to the number of clinical laboratories and healthcare providers across the US. Much of this data would need to be collected manually from the laboratories and providers. From the data sets, it would seem there is a weekly time lag.

Data Contents:

Influenza Visits – the dataset contains the number of hospital visits for patients with ILI across all US states between late-2010 and near mid-2019. Each state-year-week observation has characteristics like total ILI patients, number of providers, and total patients seen. Age categories counts have also been included but contain no data.

Lab Tests – the dataset contains the number of patients, referred to as specimens, tested for influenza. The data is organized by US state, year, and week. A percent positive variable is provided, as well as the number of positive tests broken down by influenza type (A or B) and subtypes (H1N1, H1, H3, H3N2v). Data has also been provided as to whether subtyping was not performed or not possible.

Data Limitations:

Both data sets have many missing/unreported values, represented by an X, which could limit the applicability of an analysis to certain US states. Also, some variables have inconsistently formatted variables that could be hard to interpret accurately, such as the dates within the lab tests data under the percent positive variable.

Moreover, since both data sets are based on survey data, the completeness of the data should be questioned. It may be possible that the distribution of clinical laboratories and health care providers contributing to the study does not accurately reflect influenza activity across the US. However, the CDC may have implemented counter measures for this.

Finally, the recency of the data could also be seen as a limitation given that the influenza visits data only extends to 2019 and the lab tests data extends to 2015 only. Thus, additional caution should be taken when applying any analysis from these datasets to predict future influenza activity.

Data Relevance:

I believe the influenza visits data set being of use to the project because it provides the number of patients visiting the hospital with ILIs. As this is broken up by state and week of the year, it can help the analysis identify when flu season is, its duration, and differences between states. This will help the staffing agency to understand how many additional staff is required throughout the US.

I don’t see how the lab tests data will be useful in addressing the projective objectives, so it will not be used. Both data are also not relevant to the hypothesis as age variables are absent in both data sets (or have no data).

3. CDC – Survey of Flu Shot Rates in Children

Data Sourcing:

This is external data which has been collected by The University of Chicago and sent to the CDC. Given the CDCs reputable status in addition to health providers verification of the data, this data set should be considered trustworthy.

Data Collection:

The data is collected through the National Immunization Survey (NIS) which is conducted manually via telephone interviews. The survey is done with a random sampling of parents across all US states and territories. A time lag between collection and publication of the data set is likely given the nature of labour-intensive surveys.

Data Contents:

This data contains the answers for the NIS Survey. There are many variables that mostly relate to the child's vaccination status or demographic profile. Some variables include the child's age, sex, state of residence, and vaccinations received.

Data Limitations:

Some possible limitations of data collected during the telephone interviews include the interviewees' willingness to share personal and potentially sensitive information regarding their child and self. Hence, many variables may have missing data or could be difficult to analyse accurately.

Moreover, the nature of random samples implies that generalisations made from this data may not be representative for wider populations. However, the CDC and University of Chicago seem to mitigate this by sampling parents in all US states.

Data Relevance:

Despite the large amount of information held within this data set, I think it will not be useful for the project objectives or testing the hypothesis. It's important to acknowledge that flu shots are important for managing the detrimental effects of influenza on a countries' citizens. However, the project objective is to help plan medical staffing needs across the country for flu season, not to plan or manage flu shot uptake. It may be possible to gleam insights into the affect children being vaccinated has on hospitalisation rates by state, but the hospitalisation rates are simply more valuable for determining staffing needs during this time. If the scope of this project changed to provide health centres with additional staff to administer flu shots before influenza season, this might be valuable data to have.

Also, since this data focuses exclusively on flu shots in children, this dataset can't contribute to the hypothesis about vulnerable citizens over 65.