Preparing for Influenza Season - Interim Report

Summary

An analysis is to be carried on the effects of flu and flu-related deaths to help formulate a staffing plan to provide hospitals and clinics across America with the correct staffing levels to cater for the increase in demand during Flu Season.

Project Overview

Motivation

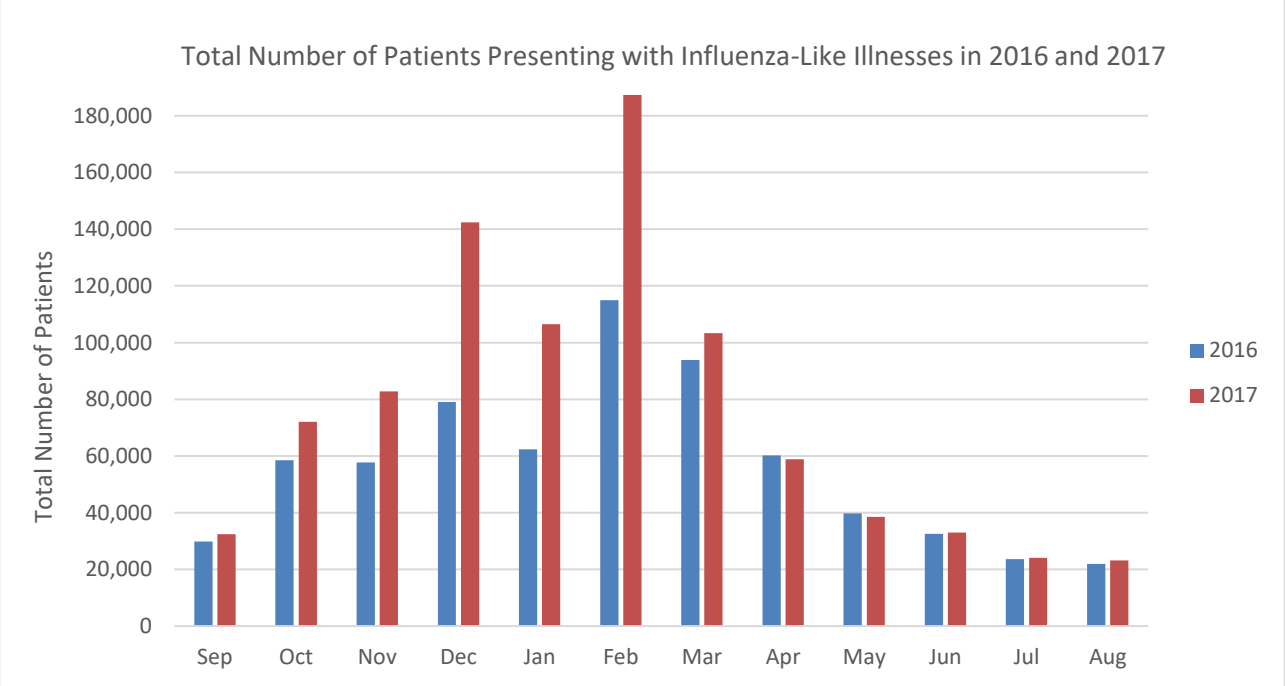
High levels of hospitalisation have been associated with influenza, particularly for those belonging to vulnerable populations, resulting in a high demand for staff. The Medical Staffing Agency are able to provide temporary staff where needed.

Objective

To devise a staffing plan to provide hospitals across America with the required levels of staffing at the most appropriate time.

Scope

The staffing plan needs to be in place for the upcoming Flu Season. Cases of patients presenting with Influenza-Like Illnesses (ILI) begin to rise as the months become cooler around October, increasing over the Christmas period, and normally peaking around February, before tailing off as the weather warms.



Research Hypothesis

"Regions with a high proportion of vulnerable individuals, are subject to higher rates of hospitalisations associated with influenza."

Refer to **Appendix A – Forming the Hypothesis** for a breakdown of thought process.

Preparing for Influenza Season - Interim Report

Data Sets

- Data Set 1: Influenza Deaths by Geography in America from 2009 to 2017.
- Data Set 2: Population Data by Geography in America from 2009 to 2017.
- Data Set 3: Influenza Survey Test Results by State - Influenza Visits from 2010 to 2019

Data Set Overview & Limitations

Data Set 1: Influenza Deaths by Geography in America from 2009 to 2017.

The data is an external source provided by the Centres for Disease Control and Prevention (CDC) through their National Centre for Health Statistics (NCHS).

The data contains monthly death counts for influenza-related deaths in the United States from 2009 to 2017. Counts are broken down further by age group, helping to highlight mortality rate figures for individuals belonging to vulnerable age groups (under 5 years & 65 years and above). However, where the death rate is recorded as <=9 (including zero), the data value is given as "suppressed", (this is in-line with data protection) but as much as 80% of the data set is affected by this value, particularly the lower age groups. Therefore, an element of imputation is required to prevent the “suppressed” values from being treated as zero, i.e., no deaths recorded.

Data Set 2: Population Data by Geography in America from 2009 to 2017.

The data set is an external source provided by the US Census Bureau. It contains population figures for 2,588 counties over 51 States across America. As well as total population, the data also provides a break down of population by gender and age group.

Census is conducted by survey, and whilst it is required to be completed by law, the data could be subject to inaccurate answers and input error. That said, this method of population count has been in operation in America since 1790 which implies that the method is fit-for-purpose.

Data Set 3: Influenza Survey Test Results by State - Influenza Visits from 2010 to 2019

The data set is an external source provided by the Centres for Disease Control and Prevention (CDC) and is collected for public health. The data set provides information on the total number of patient visits to healthcare providers in each State, in each week by year from 2010 to 2019. Of the patients that visited, the number of patients displaying Influenza-like Illness (ILI) is recorded.

The information provided by these bodies is voluntary which means that there is potential for it to not show a complete picture. However as professional bodies, all with the same goal of managing influenza, the information gathered has value and purpose.

Refer to **Appendix B – Data Set Information** for a more detailed description of the data sets used.

Note: All data sets have been trimmed by year range from 2011 to 2017.

Preparing for Influenza Season - Interim Report

Descriptive Analysis

- The end goal is to provide adequate staffing for hospitals and clinics within the various States of America during Flu Season.
- The historical data shows us:
  - The busiest months in hospitals and clinics associated with Flu Season.
  - Which regions report the most cases of ILI symptoms.
  - The death rates across the various regions, and what proportion belong to vulnerable groups.
  - The population figures for the various regions, and the proportion from vulnerable groups.
  - The approximate number of hospitals & clinics in each region, as per their participation in FluView.

The data show that there is a high proportion of individuals from vulnerable groups contributing to the mortality rate. Identifying the regions that contain a high number of vulnerable populations means that these areas can be prioritized for staffing allocation.

Core Variable Data Values Summary

Variable	Mean Value	Standard Deviation	Coefficient of Variation (CV)*
Total No. of Deaths	1,340	1,102	0.82
Total No. of Deaths from Vulnerable Groups	989	987	1.00
% of Deaths of Vulnerable Groups	66.4%	12.6%	0.19
Total Population	5,993,914	6,857,877	1.14
Vulnerable Population Count	1,205,565	1,345,286	1.12
% of Vulnerable population	20.3%	1.4%	0.07
ILI Total	13,484	21,545	1.60
No. of Hospitals & Clinics	1,684	1,682	1.00
% of Patients with ILI	1.7%	1%	0.59

*Where CV >= 1 indicates a high variation in values & CV < 1 indicates a low variation in data values.

Statistical Hypotheses

The assumption is that there is a high proportion of vulnerable individuals who die as a result of ILI. This makes the Alternative Hypothesis as follows:

H_A : The mortality rate for individuals belonging to vulnerable groups is greater than the mortality rate for non-vulnerable individuals.

$$H_A : \mu_{MortalityVulnerable} > \mu_{MortalityNon\text{-}Vulnerable}$$

Preparing for Influenza Season - Interim Report

Conversely, the Null Hypothesis is:

> *H_0 : The mortality rate for individuals belonging to vulnerable groups is less than or equal to the mortality rate for non-vulnerable individuals.*
>
> $H_0 : \mu_{MortalityVulnerable} <= \mu_{MortalityNon-Vulnerable}$

The statistical test between the two groups: *Vulnerable Individuals* and *Non-vulnerable Individuals*, was carried out using a Significance Level (α) of 0.05 (5%). The resulting P-value for this one-tail test was given as 6.235×10^{-29} which is substantially lower than the Significance Level, resulting in a rejection of the Null Hypothesis concluding that with 95% Confidence (c), the mortality rate for individuals belonging to vulnerable groups is greater that the mortality rate for non-vulnerable individuals.

Remaining Analysis and Next Steps

Analyses still to be carried out include the following:

1. Identifying which states are subject to high cases of ILI as well as high mortality rates.
2. Identify which States, and in which months, providers will be expecting a high influx of patient visits, and what proportion present with ILI symptoms.
3. Prioritize State requirements based on items 1 & 2 above as well as the estimated number of hospitals and clinics in said State.
4. Formulate a staffing allocation plan.

Remaining Project Deliverables

Week Ending	Details	Status
14-03-2021	Project plan, forming a hypothesis and initial data set review.	Complete
21-03-2021	Data set profiling.	Complete
28-03-2021	Data set comparison & statistical analysis.	Complete
04-04-2021	Hypothesis testing.	Complete
11-04-2021	Dashboard requirement plan.	Pending
18-04-2021	Tableau dashboarding for comparison & forecasting.	Pending
25-04-2021	Statistical visualization including histograms, box plots, scatter plots & bubble charts.	Pending
02-05-2021	Spatial and Textual analysis.	Pending
09-05-2021	Final presentation, findings and recommendations.	Pending

Appendices

Appendix A – Forming the Hypothesis

Appendix B – Data Set Information

Appendix A - Forming the Hypothesis

Chosen Hypothesis

"Regions with a high proportion of vulnerable individuals, are subject to higher rates of hospitalisations associated with influenza."

What is a Hypothesis?

A hypothesis is a prediction or assumption, and involves the relationship between two variables: the ***Independent*** variable, (the variable that is changed as part of the test), and the ***Dependent*** variable, (the variable that is observed for changes based on changes in the Independent variable).

What Data is Available?

1. Death rates associated with influenza.
   - By State, Year, Month, Age Group.
2. Population rates.
   - By County & State, Gender, Age Group.
3. Influenza laboratory test results.
   - State, Week No (month), Total Visits, Total Providers, Total ILI.
4. Survey of flu shots in children.
   - By State, Year, Age in months, plus demographic information.

Considering the Variables

Who suffers the most with the flu?	– Vulnerable groups
When do they suffer?	– Cold season mostly
Where do they suffer?	– States with high levels of vulnerable populations.
What is the root cause of the problem?	– Flu & death from flu
What is a contributing factor?	– Understaffing & vulnerable populations
What helps to alleviate the issue?	– Appropriate staffing level allocation to populations with historically high rates of death-by-flu and/or with large numbers of vulnerable populations.

Appendix B - Data Set Information

Data Set 1

Influenza Deaths by Geography in America from 2009 to 2017.

Data Source

The data set is an external source provided by the Centres for Disease Control and Prevention (CDC) through their National Centre for Health Statistics (NCHS). Government data is considered a trustworthy source.

Data Collection & Potential Limitations

The data is administrative data collected as part of the National Vital Statistics Cooperative Program. Each of the U.S. states and territories is required to record all births, deaths, marriages, and divorces within their jurisdiction. Death records come from death certificates, in which a doctor codes the primary cause of death as “Influenza” or “Pneumonia” (ICD-10 codes J09-J18).

Because this data is part of the government's vital statistics program, and similar to a census, the data is assumed a complete and accurate count of deaths.

One caveat, however, is that deaths on a death certificate only list one cause of death. This could create some discrepancies within vulnerable populations, such as those with AIDs—while the cause of death may be related to AIDs, their decline in health may have been initiated by influenza.

Data Contents

The data contains monthly death counts for influenza-related deaths in the United States from 2009 to 2017. Counts are broken down further by age group, helping to highlight mortality rate figures for individuals belonging to vulnerable groups. That said, where the death rate is recorded as <=9 (including zero), the data value is given as "suppressed". This is in-line with data protection. However, as much as 80% of the data set is affected by this value, particularly the lower age groups. Therefore, an element of imputation is required to prevent the “suppressed” values from being treated as zero, i.e., no deaths recorded.

Data Relevance

The data shows the geographic and monthly spread of influenza across the United States over multiple years. With historical trends often mirror upcoming trends, the historical influenza data can be used to predict future influenza seasons for planning purposes.

Appendix B - Data Set Information

Data Set 2

Population Data by Geography in America from 2009 to 2017.

Data Source

The data set is an external source provided by the US Census Bureau. It is publicly owned and because this is government data the data set can be considered viable.

Data Collection & Potential Limitations

The US Census Bureau conducts a population count on 1st April every 10 years on years ending in zero. The population for subsequent years is then estimated using a variety of sources and methods.

Census is conducted by survey, and whilst it is required to be completed by law, the data could be subject to inaccurate answers and input error.

Population estimates use other data sources of an administrative nature such as data provided by the National Centre for Health Statistics (NCHS) on birth, death and migration rates.

The last census, in the provided data set, was carried out in 2010 meaning there have been seven years of estimation since. The way in which population is counted has been operational in the US since 1790. This implies that the methods used, i.e., a census every 10 years followed by informed estimates for subsequent years, is a successful indication of population count.

Data Contents

The data set contains population information by county split into the following groups:

- Overall population
- Female population
- Male population
- Population by age range, split into 17 groups of five-year ranges plus an 'over 85 years' group.

Data Relevance

The data set is highly relevant to the project for the following reasons:

1. Population count per State can help to calculated the % effected by death in each State, and therefore how severely the community are affected by influenza.
2. Population counts for vulnerable age groups (under 5 and over 65) can help to calculate the % of the population that are assigned to a vulnerable group owing to age.

Appendix B - Data Set Information

Data Set 3

Influenza Survey Test Results by State - Influenza Visits from 2010 to 2019

Data Source

The data set is an external source provided by the Centres for Disease Control and Prevention (CDC). The data is collected for the public health, and is publicly owned. For these reasons, it is considered reliable.

Data Collection & Potential Limitations

The CDC runs as a collaborative surveillance program called FluView where weekly data is collected, compiled and analysed on influenza activity. As well the CDC, the bodies that contribute to the data used for FluView include:

- State, local, and territorial health departments,
- Public health and clinical laboratories,
- Vital statistics offices, such as the National Centre for Health Statistics.
- Healthcare providers & Clinics,
- Emergency departments.

The information provided by these bodies is voluntary which means that there is potential for it to not show a complete picture. However as professional bodies, all with the same goal of managing influenza, the information gathered has value and purpose.

NB: The original data provides a breakdown of Age Groups but there is no data reported for these columns.

Data Contents

The data set provides information on the total number of patient visits to healthcare providers in each State, in each week by year from 2010 to 2019. Of the patients that visited, the number of patients displaying Influenza-like Illness (ILI) is recorded.

Data Relevance

The data set is of relevance to the project as we can calculate the % of patients affected by ILI and compare that to population.