Preparing for Influenza Season - Interim Report

Sarah Tischer

tl;dr: *The analysis aims to optimize medical staffing during the U.S. influenza season, with a focus on vulnerable age groups. Initial analysis indicates a weak correlation between population size and influenza deaths in the 65+ age group. Further investigation supports the hypothesis that individuals over the age of 65 face a higher influenza mortality rate, suggesting that in-depth analysis, visualization, and qualitative insights are needed as next steps to inform effective staffing plans.*

Project overview

Motivation	*The United States has an influenza season where more people than usual suffer from the flu. Some people, particularly those in vulnerable populations, develop serious complications and end up in the hospital. Hospitals and clinics need additional staff to adequately treat these extra patients. The medical staffing agency provides this temporary staff.*
Objective	*Determine when to send staff, and how many, to each state.*
Scope	*The agency covers all hospitals in each of the 50 states of the United States, and the project will plan for the upcoming influenza season.*

Research hypothesis

The age groups 0-5 and 65+ develop more severe flu complications. If you are under 5 or over 65 years old, then you have a higher risk of dying from the flu.

Data overview

Population data by geography, time, age, and gender

This data set from the US Census Bureau contains population statistics by county from 2009 to 2017. The data is categorized by gender and various age groups.

Influenza deaths by geography

This data set from the CDC contains monthly counts of influenza-related deaths in the United States from 2009 to 2017. The counts are divided into two categories: state and age.

Data limitations

Population data by geography, time, age, and gender

- Accuracy: The numbers in each category are estimates and may not be completely accurate. In addition, because the data was collected manually through surveys, it is susceptible to human error (e.g., incomplete information, transmission errors).
- Data biases: The census relies on sampling techniques, and if the sample is not representative of the entire population, it can introduce bias. Additionally, the response rate to surveys can vary, and a low response rate may introduce non-response bias.

Influenza deaths by geography

- Accuracy: The data source is based on death certificates in which providers report only a single underlying cause of death, which could lead to masking and inaccurate reporting of patients with pre-existing conditions.
- Data suppression: Statistics representing less than ten people (0-9) are suppressed to protect confidentiality.

Descriptive analysis

The research hypothesis proposes that a higher population in the 65+ age group will result in an overall higher number of influenza deaths in the same age group.
A descriptive analysis was conducted to assess the quality of the data and to understand the similarity and variance between the variables. Therefore, the mean and standard deviation were calculated for both the number of influenza deaths over 65 years and the population over 65 years.

	Influenza deaths over 65 years	Population over 65 years
Mean	*826*	*806989*
Standard Deviation	*1014*	*887017*

The relatively high standard deviation for influenza deaths suggests that the number of influenza deaths in this population varies significantly from year to year and/or from state to state.
To further determine if there is a relationship between the two variables, they were correlated.

- Correlation coefficient: ***0.049067792***
- Strength of correlation: ***weak to no relationship***

The correlation coefficient weakly supports the hypothesis that individuals aged 65 and above are more prone to severe influenza complications and a higher risk of death. Instead, factors like vaccination rates, access to healthcare, general health, and various demographic and environmental elements seem to influence influenza deaths in this age group.
In conclusion, the relationship between population size and influenza deaths in the 65+ age group isn't particularly informative for explaining variations in the hypothesis variable. Further in-depth analysis is crucial before drawing conclusions.

Results & insights

To gain further insight from the sample data sets, inferential statistics were used to translate the research hypothesis into conclusions about the population as a whole.

Null hypothesis

People > 65 years have the same or lower influenza mortality rate as people < 65.

Alternative hypothesis

People > 65 years have a higher influenza mortality rate than people < 65.

t-Test: Two-Sample Assuming Unequal Variances		
	Influenza mortality rate of people younger than 65 years	*Influenza mortality rate of people 65 years and older*
Mean	78.76470588	826.2875817
Variance	22903.91395	1028483.747
Observations	459	459
Hypothesized Mean Difference	0	
df	478	
t Stat	-15.61886217	
P(T<=t) one-tail	4.95504E-45	
t Critical one-tail	1.648047653	
P(T<=t) two-tail	9.91009E-45	
t Critical two-tail	1.964939272	

The test shows that the p-value (i.e., the probability that the difference in mortality rates between the age groups is due to chance) is well below the specified significance level of 0.05. This allows us to reject the null hypothesis with 95% confidence.

In other words, there is a 95% probability that the mortality rate is affected by age, and therefore people 65 and older are more likely to die from the flu than people younger than 65.

Remaining analysis & next steps

Remaining analysis

Having determined that this age group is a vulnerable population, the following steps will be conducted to fully identify the variables that affect influenza mortality and to provide a compelling case for prioritization across states.

Temporal Visualizations	*Determine if there are differences in the mortality rates for this age group over time to identify previous interventions that may have had an effect (e.g., vaccination promotion, community health programs).*
Statistical Visualizations	*Categorize each U.S. state as low, medium, or high need, based on the number of vulnerable populations, so that high-need states can be prioritized.*

Spatial Analysis	*Determine if there are differences in mortality rates for this age group by geographic region, recognizing that external variables (e.g., access to health care, climate) may also affect influenza seasonality and/or mortality rates.*
Textual Analysis	*Gather qualitative insights such as * perceptions of the effectiveness of staffing plans * patient sentiment about experiences at particular hospitals * vaccination concerns*

Next steps

Upon completion of the analysis, a video presentation will be recorded to present the findings to stakeholders.

Appendix

'Project Overview'

Business Requirements Document

'Hypothesis'

Clarifying questions

1. Which group from the vulnerable population most frequently ends up in hospital?
2. What is the required patient-to-staff ratio?
3. Which states have the highest flu shot rate?

Funneling questions

Which group from the vulnerable population most frequently ends up in hospital?

1. Do the same groups of vulnerable populations end up in hospitals in every state?
2. How long on average does a patient with influenza have to stay in the hospital?
3. Is the mortality rate of influenza patients the same for each group?
4. Are there any thresholds or criteria for determining the level of vulnerability?

What is the required patient-to-staff ratio?

1. Are any states currently over- or understaffed?
2. Do any states have public health efforts supporting influenza prevention?

Which states have the highest flu shot rate?

1. How are vaccinations organized across different states?
2. Are there any vaccination restrictions for vulnerable populations?
3. How long does it take for a vaccination to protect against infection effectively?

<u>Privacy and ethical questions</u>

1. How should sensitive patient data be handled and protected, especially when dealing with information related to vulnerable populations? What measures are in place to ensure compliance with healthcare data privacy regulations such as HIPAA (Health Insurance Portability and Accountability Act)?

2. How should potential biases in the data be addressed, particularly in relation to vulnerable populations? Are there strategies in place to mitigate bias and ensure fair representation, considering that certain demographic groups may be disproportionately affected by influenza?

'Data Overview'

Population data by geography, time, age, and gender

<u>Data sourcing:</u>

This is an external data source, provided by the US Census Bureau. As government data, it can be verified as a trustworthy data source.

<u>Data collection:</u>

The data is administrative data manually collected by the US Census Bureau via surveys conducted by mail, over the phone, in person, online, or some combination of these. The census is conducted periodically, and demographic characteristics of the population can change between census years. This can lead to inaccuracies if there are significant demographic shifts during the intercensal period.

<u>Data contents:</u>

The data set contains population statistics by county, broken down into categories for gender and different age groups, from 2009 to 2017.

<u>Data limitations:</u>

- The numbers in each category are estimates and may, hence, not be completely accurate.

- Since the data has been collected manually via surveys, they are prone to human errors (e.g. incomplete information, transmission errors).
- The response rate to surveys can vary, and a low response rate may introduce non-response bias.
- The census relies on sampling techniques, and if the sample is not representative of the entire population, it can introduce bias.

Data relevance:

The data set contains crucial information about the distribution of age groups in the different states. This is particularly important for determining where most people in the relevant age groups 0-4 and 65+ live. As it was collected via the US Census Bureau, it can be assumed that it's the most trustworthy and complete version of the data available and will, hence, be used for the analysis.

Influenza deaths by geography

Data sourcing:

This is an external data source. The medical staffing agency doesn't have this information, so it's relying on government data. The data is provided by the Centers for Disease Control and Prevention (CDC) through their National Center for Health Statistics. As government data, it can be verified as a trustworthy data source.

Data collection:

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, it's similar to a census, meaning that it can be assumed a complete and accurate count of deaths. The 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 into two categories: state and age.

Data limitations:

- The data source is based on death certificates in which providers report only a single underlying cause of death, which could lead to masking and inaccurate reporting of patients with pre-existing conditions.
- Statistics representing less than ten people (0-9) are suppressed to protect confidentiality.

Data relevance:

The data shows the geographic and monthly spread of influenza across the United States over multiple years. As it was collected via the government vital statistics program, it can be assumed that it's the most trustworthy and complete version of the data available.

Historical trends often mirror upcoming trends. For this reason, the historical influenza data can be used to predict future influenza seasons for planning purposes. The project objective asks when and where to send staff, which is something this data can address. Additionally, one of the project requirements is to prioritize vulnerable populations. The included demographic data can help illuminate vulnerable populations that require additional care when planning for influenza. For these reasons, this data set is critical to addressing the project objective.