NW Data Science Boot Camp

PROJECT 1 - August 2021

Katharine Hollars, Nick Prieto, Rachel Verner, Steve Reiss

**PROJECT**

COVID has focused greater attention on public health issues, especially related to cause of death. We were interested in exploring causes of death related to disease - both how they were impacted by a variety of demographic factors and how they were impacted by the onset of COVID.

**DATA SOURCE**

https://data.cdc.gov/NCHS/AH-Monthly-Provisional-Counts-of-Deaths-by-Age-Gro/ezfr-g6hf which we downloaded and saved as “death\_by\_disease.csv”

Fields include:

* REGION: Department of Health and Human Services (HHS) regions (1 - 10) and US total
* AGE: Predefined age ranges
* DISEASE: Death by disease type
* DATE: Month and year (January, 2019 - March, 2021)

**HYPOTHESIS**

Cause of death by disease varies by region, age, and season. The emergence of COVID significantly impacted the distribution of death by disease.

**QUESTIONS**

1. REGION: What are the top three causes of death by HHS region? DId the onset of COVID change this, and how?
2. AGE: Is age a factor in the cause of death by certain diseases? How do specific diseases have different death count distributions by age group?
3. SEASONALITY: Does seasonality play a role in the distribution of death counts? What does the seasonality of COVID look like compared to other respiratory diseases?

**CLEANING THE DATA**

* Simplifying our data:
  + Cleaned up data so it was easier to review and sort - removed unnecessary columns, converted NaN values to zero, renamed columns for clarity
* Creating regional breakouts:
  + Created a data frame for US Total, and breakouts for each HHS region
  + NOTE: HHS data reflects regional totals only - some regions have significantly larger populations than others so we only examined regional trends, not per capita differences between the regions
* Identifying key segments:
  + Grouped by both age and disease type for future manipulation
* Unanticipated data issues:
  + Data had been pre-cleaned and needed to be re-indexed to display data correctly, no data dictionary was provided - so only included columns with clear definitions, data did not include all diseases, only the most prevalent causes of death

**ANALYSIS**

***Question 1: death by region***

* COVID deaths by region:
  + Over our date range (results may have changed since March, 2021) trends were consistent by region, increasing with age
* Top three causes of death by region:
  + In 2019 the top three causes of death were fairly consistent between regions - with most deaths attributed to heart disease and malignant neoplasms (cancer)
  + In 2020 COVID (multiple cause of death) became a top three cause of death in all regions
  + In region 2 COVID (underlying cause of death) and COVID (multiple cause of death) were both top three causes of death. This may be specifically related to special conditions in New York City

***Question 2: death by age***

* Heart disease and cancer:
  + Heart disease and cancer are the leading causes of death fairly equally under age 55
  + From age 55 to 75 cancer becomes the leading cause of death
  + Over age 75 and increasing dramatically over 85, heart disease becomes the leading cause of death
* Covid deaths - underlying vs. multiple causes:
  + We found no difference between COVID (underlying cause of death) and COVID (multiple cause of death) across different age groups
* Children:
  + Children age 0-4 much more likely to have a natural cause of death than children 5-14
  + Children age 0-4 had death by disease distributed between several primary causes, in children 5-14 it was overwhelmingly related to cancer

**Question 3: death by seasonality**

* Trends:
  + Overall death by disease counts were higher in the winter months but were also higher in April
* Respiratory disease:
  + Death by respiratory disease (both “influenza pneumonia” and “other diseases of the respiratory system”) show winter seasonal increases but did not account for the increase in April
* COVID:
  + COVID deaths spiked in April as well as during the winter months, contributing to the higher totals at both times
  + Unlike respiratory diseases, there is no seasonal pattern for COVID deaths

**ADDITIONAL QUESTIONS, NEXT STEPS**

* Our analysis supports our hypothesis that cause of death by disease varies by region, age, and season and that the emergence of COVID significantly impacted the distribution of death by disease.
* Further analysis based on our findings might include:
  + Compare the HHS regions on a per capita basis to expand our analysis beyond trends to reflect actual death rates
  + Analyze regional data using weather reports to see how regional weather differences impacts cause of death and seasonality
  + Look at more recent COVID data to examine the impact of COVID since vaccines became available, especially by region given varying vaccination rates
  + Examining a wider range of diseases, based on access to an expanded cause of death dictionary
  + Look more closely at distribution of death by heart disease and cancer, taking into account the additional information above

**RELATED RESOURCES**

* Presentation: [*https://docs.google.com/presentation/d/1yRLU\_pJdZdACNOV0aKRaX7uTJcpZP4dcSXAVitexjmg/edit#slide=id.gc6f980f91\_0\_81*](https://docs.google.com/presentation/d/1yRLU_pJdZdACNOV0aKRaX7uTJcpZP4dcSXAVitexjmg/edit#slide=id.gc6f980f91_0_81)
* GitHub: [*https://github.com/shindokl90-lab/Project-One.git*](https://github.com/shindokl90-lab/Project-One.git)
* Project Summary (this document): [*https://docs.google.com/document/d/1ufye6WOh16TxEBMU\_OcWJxQHMcswN3A\_vcFHdlNYc1I/edit#*](https://docs.google.com/document/d/1ufye6WOh16TxEBMU_OcWJxQHMcswN3A_vcFHdlNYc1I/edit)