DATA 512 Project Part 2 – An Extension Plan

(Analysis of Google Search Patterns during the COVID-19 pandemic)

Motivation/Problem Statement

During any pandemic, individuals seek health information due to media curiosity, their own illness, or the illness of a family member. Health information seeking, also known as infodemiology, has been proposed as another disease surveillance method. During the epidemic, digital media played a massive role in spreading information. In addition to conventional surveillance, digital surveillance (internet) can provide intriguing trends concerning COVID-19's general concerns and aid in improving prediction models. The present digital surveillance analysis is driven by individuals' worries in the face of a pandemic. It is being used to determine the extent to which the Public Mask Mandate can safeguard people from getting COVID-19 symptoms. Google Trends (GT) is a valuable resource that may be used for digital monitoring. It may predict healthcare decisions in real-time by following people's search behaviors and recording their worries (symptoms, vaccination availability, eligibility, and side effects).

Milwaukee is one of Wisconsin's largest counties with the highest number of detected COVID-19 cases. Illness-specific symptom search patterns on GT can alert the county's healthcare system to prepare and allocate resources needed ahead of time. The primary goal of this analysis is to leverage the GT to understand the trends in Milwaukee County citizens' search patterns for symptoms and correlate this information with the number of cases, fatalities, vaccinations, and impact of the public mask mandate in Milwaukee County. The findings can substantially aid in estimating the need for supplementary monitoring and policy methods and informing real-time public health choices.

Also, Vaccine hesitancy remains a severe challenge in ending the COVID-19 pandemic. Unfortunately, online platforms can also spread substantial misinformation about vaccines. The secondary goal of the analysis is to leverage the GT to understand the trends in vaccine misinformation and how public acceptance of vaccines changed over time with changes in mask mandate policies. Using this analysis, we can attribute these trends to public attitudes, interests during the varying vaccine availability, misinformation regarding the vaccines leading to a high unvaccinated population, and further target education among the citizens.

The proposed analysis can make the policy changes more human-centered as it indirectly incorporates the participatory design strategy. The epidemiological information about the virus spread in the county alone isn't sufficient to devise an effective plan to attenuate the infection rate and accelerate vaccine administration. Embedding the end users' thoughts and concerns about the pandemic and vaccinations can provide real-time, immediate, quick, and cheap feedback to the algorithms that predict/decide policy changes. For example, local governments can develop effective vaccine distribution algorithms and make accurate infection predictions if they know the concerns of the citizens of the county. It eliminates the scope for wrong interpretations and provides more transparency to the end users of the algorithms developed specifically for the county. Uncovering these wide ranges of insights while considering the context of virus spread reveals insights about individual behaviors and the relationships between different entities involved in the policy-making process.

Research questions and Hypotheses

The data sets collectively are powerful in answering many research questions related to the COVID-19 pandemic. However, I want to limit the scope of my analysis to associations with masking mandates. Below are a few research questions I am planning to answer using the analysis.

Note: The hypothesis mentioned below are examples, and the numbers and names shouldn't be taken at face value

Research Question 1 – How does the Public Mask Mandate in the county impacts the search for all COVID-19 symptoms?

Hypothesis 1 –

- Individuals are X% less likely to search for symptoms if the Public Mask Mandate is effective in the county.
- The mask mandate decreases the likelihood of searching for any symptoms within the last X days by Y%.

Research Question 2 – What are the top symptoms strongly impacted by the mask mandate?

Hypothesis 2 – People are X%, Y%, and Z% less likely to search for fever, chills, and fatigue in Milwaukee County during Mask Mandate

Research Question 3 - What are the highly correlated symptom search terms with the daily confirmed COVID-19 cases and fatalities before and after the mask mandate policy?

Hypothesis 3 –

- **"**Severe chest pain" positively correlates with COVID-19 fatalities when no mask mandate exists.
- A time-lag correlation of X days is observed between the daily number of confirmed cases with the top symptom search terms "cough" and "shortness of breath" during the mask mandate.

Research Question 4 - How are vaccination intent and side effect search terms correlated with mask mandate policies?

Hypothesis 4 – Individuals showed X% more vaccination intention when the masking mandate was removed in the county

Data to be used

Additional to the epidemiological data used in part 1 of the analysis, I plan to use the below search trends data sets to analyze the digital surveillance of COVID-19 in Milwaukee county. This additional data will help us understand how the progression of the virus is associated with growing concerns and sentiments among people and how the public mask mandate significantly lowered the incidence of developing all/top COVID-19 symptoms.

Data Source 1:

- Data set - COVID-19 Search Trends symptoms dataset
- Link - https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-search-trends.md
- Description – The dataset consists of aggregated, anonymized trends in Google searches for more than 400 health symptoms, signs, and conditions, such as cough, fever, and difficulty breathing. The dataset provides a time series for each region, showing the relative volume of searches for each symptom

Data Source 2:

- Data set - COVID-19 Vaccination Search Insights
- Link - https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-vaccination-search-insights.md
- Description - This aggregated, anonymized data shows trends in search patterns related to COVID-19 vaccination. These trends in search patterns are made available with the intention of helping design, target, and evaluate public education campaigns. These trends reflect the relative interest in Google searches related to COVID-19 vaccination.

Terms of use – The GT data used for this analysis uses Google Terms of Service.

Ethical Considerations – For the above two datasets, differential privacy has been used by adding artificial noise, enabling high-quality results without identifying anyone. To further protect people's privacy, it is ensured that no personal information or individual search queries are included in the dataset.

The above datasets will have the RSV (relative search values) of the symptom and vaccination search terms. Analyzing the trends using these values and the mask mandate information will provide evidence for the health benefits of wearing masks in public in the initial stage of the COVID-19 pandemic. The vaccination search trends will help highlight the relevance of public mask-wearing for the ongoing pandemic when the concerns about the side effects of vaccines are more significant in the county.

Unknowns and Dependencies

The data collected limits the searches made by people on the Google platform. This data might be skewed with searches from one platform, varying connectivity, and only people who agreed to Google's terms of use. When the data doesn't meet quality and privacy thresholds, we might see empty fields for specific dates.

Additionally, a few dependencies come from using the data set because of other factors happening in the same timeline. For example, symptom searches can happen due to multiple other factors like Flu and other seasonal, regional illness. Also, the vaccine side effect concerns could've been raised because of traditional and political thought processes. Hence, assumptions are being made for the analysis by keeping other external factors constant.

Methodology

The analysis methods I plan to perform are broadly classified into two techniques – Regression and Correlation Analysis and supporting statistical tests to validate the underlying assumptions. Using these techniques, I plan to draw meaningful inferences about the trends in digital surveillance, given the changing virus dynamics and mask policies.

Regression Analysis:

Analysis 1: Association between Symptom searches and Infection rate
The relationship between the numbers of daily cases and fatalities with RSV of the symptom searches can be modeled with regression analysis. Below are the predictor and response variables that will be individually fitted to the regression model.

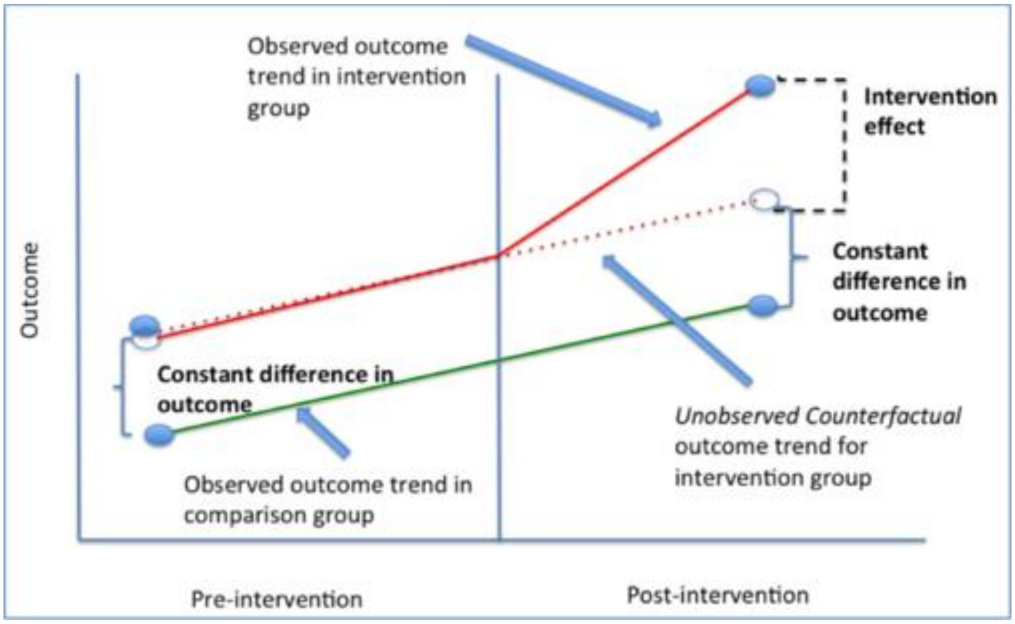
Response Variables (Y) – Number of Confirmed Cases, Number of Fatalities
Predictor Variables (X) – Symptom search term RSV values (Cold, Fever, Chest Pain, etc.)

The coefficient values associated with the fitted regression model will help us understand the impact of one variable on another. This will help us quantify the changes in people's searches with the progression of the infection.

Based on the validity of the regression model assumptions, the respective methods will be used to identify the coefficients of the symptom RSV variables. As the data is time series and can have non-normal residuals and time-dependent variables, the below methods will be selected accordingly.

Ordinary Least Squares (OLS) – Homoscedasticity, Normally distributed
Weighted/Generalized Least Squared (WLS/GLS) – Heteroscedasticity, Non-Normal
Vector Auto Regression (VAR) – Multi-variate and time-dependent

Analysis 2: Impact of Public Mask Mandate on Symptom search terms
I will use the Difference-in-Difference (DID) regression technique to model the differential timing of the mask mandate implementation in Milwaukee County. In this DID framework, we compare the health outcomes for individuals under the Public Mask Mandate period with those in the same county who were reported when the mandate had not been enforced. We can also examine the impacts of the Public Mask Mandate for each symptom individually. Below is the visual representation of the observable effect using the algorithm.


Observed outcome trend in intervention group
Intervention effect
Constant difference in outcome
Outcome
Constant difference in outcome
Unobserved Counterfactual outcome trend for intervention group
Observed outcome trend in comparison group
Pre-intervention
Post-intervention

The coefficient of interest β1 summarizes how the Public Mask Mandate affects individuals' search for COVID-19 symptoms. The symptom searches with the highest coefficient values quantify the highest impacted symptoms due to the masking policy.

Statistical Tests:
Chi-Squared (Test of Independence) - A chi-squared test of independence can be used to check the dependence of mask mandate with confirmed cases and symptom search volume. Below is the sample table with the average RSV used to test the independent hypothesis between the variables.

Mask Mandate/Symptom Search	Cold	Throat Pain	Chest Pain

Yes	7.8	1.6	1.2
No	10.2	2.5	1.8

Shapiro-Wilk (Test for Normality) - To check the normality assumption for the regression model, we can use the Shapiro-Wilk test on each top symptom RSV variable. The significance of these tests will decide which variable coefficient values can be used for inferences.

Correlation Analysis:

Analysis 3: Temporal relationship between Symptom searches and infection rate
To understand the direction and strength of the association between the symptom-related search terms and infection rate, we can calculate the time-lagged correlation values (Pearson correlation coefficients) and *P* values for each symptom-related search term. Dynamic conditional correlation (DCC) and sliding windows correlation models will be leveraged to obtain these values for the top symptom searches. The best model with the least RMSE values will be chosen for analyzing the time-varying correlation values.

Presentation Methods:

The below visuals will be represented on the Tableau dashboard.

1. Regression plots describing the relationship between the daily cases and fatalities with the Symptom RSV. It represents how both variables are associated with high and low values of the variables.
2. Time-lagged correlation plots overlapped for the top 5 symptom searches will be visualized to show the lag days with the associated correlation values. It effectively represents the day with the best correlation between both variables (i.e., symptom keyword and confirmed COVID-19 cases).

Timeline to Completion

Below are the critical timelines and project tasks.

Nov 10, 2022 – Problem statement submission (Part 2)

Data finalization and documentation of the motivation, hypothesis, and methodology

Nov 15, 2022 – Data extraction and exploratory data analysis

Extract the data specific to the county and perform exploratory data analysis to understand the general trends in symptom search terms

Nov 18, 2022 – Regression, DID, and Correlation analysis

Extension to part 1 by analyzing important search patterns that are highly correlated with the progression of daily cases and fatalities in the county and analyze the impact of masking mandate on these symptoms using the difference-in-difference model

Nov 25, 2022 – Model diagnostics and inference analysis

Validate the model assumptions, run diagnostics and draw inferences from the analysis

Dec 2, 2022 – Data Visualization

Visualization of highly correlated search patterns and vaccination search trends

Dec 5, 2022 – Presentation of key takeaways and findings (Part 3)

Creating a presentable story with takeaways and findings from the regression and correlation analysis

Dec 12, 2022 – Final report documentation (Part 4)

Submitting the final report documenting the analysis, inferences, and any future work

References

Datasets:

[1] Google LLC "Google COVID-19 Vaccination Search Insights". http://goo.gle/covid19vaccinationinsights, Accessed: 2022-11-10.

[2] Google Open Data - https://health.google.com/covid-19/open-data/raw-data

[3] Symptom Search trends - https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-search-trends.md

[3] Vaccine Search Trends - https://github.com/GoogleCloudPlatform/covid-19-open-data/blob/main/docs/table-vaccination-search-insights.md

Current Literature:

[1] https://pair-code.github.io/covid19_symptom_dataset/?country=US

[2] Kurian SJ, Bhatti AUR, Alvi MA, Ting HH, Storlie C, Wilson PM, Shah ND, Liu H, Bydon M. Correlations Between COVID-19 Cases and Google Trends Data in the United States: A State-by-State Analysis. Mayo Clin Proc. 2020 Nov;95(11):2370-2381. doi: 10.1016/j.mayocp.2020.08.022. Epub 2020 Aug 20. PMID: 33164756; PMCID: PMC7439962.

[3] Nguyen M. Mask Mandates and COVID-19 Related Symptoms in the US. Clinicoecon Outcomes Res. 2021 Aug 16;13:757-766. doi: 10.2147/CEOR.S326728. PMID: 34429625; PMCID: PMC8379388.