**Business Understanding**

Despite being a common annual occurrence, the influenza virus, also known simply as the flu, is one of the leading causes of death in the United States and around the world. By depressing the immune system of the host, especially effectively in vulnerable children and elderly, influenza paves the way for more dangerous lung infections to cause pneumonia. In 2017 influenza was associated with 55,672 deaths in the U.S., making it the 8th leading cause of death that year. Although flu vaccines are an important tool available to reduce the public health impact of influenza, because of the slow process with which these vaccines are manufactured they have to be made months ahead of the flu season. The long lead times force manufacturers to develop the vaccine based on an estimate of the influenza virus strain that will be dominant in the upcoming season, which can result in the vaccine being less effective if the guess is wrong. Because of this high uncertainty, the Center for Disease Control (CDC) has made it a national priority to track the spread of influenza by maintaining a weekly national database of key metrics related to influenza. The two most important of these metrics are the influenza-like-illnesses (ILI), which tracks the number of patients who presented themselves to the healthcare system with symptoms of influenza, and a count of the influenza cases that were confirmed by laboratory tests. The goal of the CDC is to use this data to better predict the spread of the flu in order to make better plans about resource allocation and public engagement in the vaccination campaign with the aim of reducing the number of infections and deaths. Although these metrics are reported on a weekly basis, there is a one week lag period between hospitals collecting the data and the results being reported because the CDC requires a week to aggregate and analyze the data that they’ve received. It is a critical mission for researchers to fill this knowledge gap with an effective forecasting model because the incubation period for the flu is an average 2 days. Currently the CDC issues notices about elevated flu seasons, but if the CDC were able to forecast a surge in influenza cases in advance it would be able to issue an early warning to the healthcare system to prepare their emergency rooms as well as increase their public vaccination awareness efforts. Previously attempts have been made to improve the forecasting model with weather data, as temperature and precipitation are important to the flu incubation process. In addition, data science tools have been used to improve the model with the inclusion of search and social media data. However, many of these attempts are at the location-insensitive national level or only focus on one county. In addition, recent health policy research has shown that economic and social factors are important determinants of health. We develop a model for all NY state counties and attempt to improve the model with a unique recently-released county-level economic dataset.

**Data Understanding**

The CDC publishes its flu surveillance data and maintains an interactive visualization tool on its FluView website.[[1]](#endnote-1) The data available publicly on FluView, however, is aggregated to the state level which makes it less precise than data available at the county level. The primary database used is a county level database specific to the State of New York for the period 2009 and 2019 reporting the number of laboratory-confirmed influenza cases.[[2]](#endnote-2) The target variable is chosen to be laboratory-confirmed influenza cases over the ILI metric because only 8% of people who visit the doctor for the flu actually have the infection and the ILI contains noise from other seasonal diseases such as the common cold. The Laboratory-Confirmed Influenza Case dataset contains 20,080 county/week pairs and provides the data for the county, the weekend date, the flu season week as defined by the CDC, and the flu counts for that period broken down by the type of flu (A,B, or unspecified).

Weather is an important factor in the spread of flu because the flu spreads most effectively in cold and dry seasons.[[3]](#endnote-3) The National Oceanic and Atmospheric Administration provides local weather data through its regional climate centers (RCC). Data for New York state counties is sourced from the Northeast RCC’s Climod system. The system allows to search for a city or county and provides a map of nearby weather stations. The station chosen for each county is the station closest to the county centroid and without any significant gaps in the data. The weather data provides daily max temperature, minimum temperature, average temperature, precipitation, snowfall, and snow depth.

Google Trend provides relative search popularity for various search terms on a weekly and regional basis.[[4]](#endnote-4) Notably the Google Trend reports on a weekly basis for data pulls that contain a maximum of 5 years, and reports on a monthly basis for longer periods. Data is pulled in three pulls for periods 2009-2011, 2012-2015, and 2016-2019 for the search terms “flu”, “flu symptoms”, “fever”, “cough”, and “sore throat”.

The CDC collects and provides data on the FluVaxView site on the vaccination rate estimates from the Behavioral Risk Factor Surveillance System (BRFSS) and National Immunization Survey (NIS).[[5]](#endnote-5) The FluVaxView data contains monthly state-level estimates for vaccination rates for adults and children, but because the children data is differentiated between New York City and the rest of the state, only this data is used in the model.

Income and employment are important social determinants of health, making economic activity a novel addition to influenza modeling. Although economic data is normally reported at the state level, the Bureau for Economic Analysis has released county level GDP data for the period 2012-2015 from which GDP and GDP change data are added to the model[[6]](#endnote-6). In addition, the United States Department of Agriculture provides annual county-level demographic data in the Atlas of Rural and Small-Town America.[[7]](#endnote-7) This data source is used for population level and unemployment level estimates for the model.

**Data Preparation**

The target variable being forecast in this model is chosen to be the total number of laboratory-confirmed influenza cases in a county in two weeks from the time of the most recent case report. This means that, for example, as the CDC is publishing the final analysis for week 30 in week 31, the model is estimating the number of confirmed cases in week 32. Although this limits the amount of public health impact that can be gained from this early warning, the CDC also has a responsibility to maintain the public’s confidence in the government. Recent trends of active resistance to vaccination in the anti-vax movement could be emboldened by inaccurate CDC forecasts, and longer forecast horizons would decrease the accuracy of the model. The 2-week forecast (1-week warning) target decreases this risk and is still enough time to have some impact against the 2-day influenza incubation period, as well as allow hospitals and other providers to prepare for a surge.

Before the datasets could be merged many had to be reformatted to fit the county/week # index used in the primary Laboratory-Confirmed Influenza Cases dataset.

The Laboratory-Confirmed Influenza dataset reports each county/week pair in three lines corresponding to the type of influenza that was confirmed (A, B, unspecified). These lines are merged to obtain a single aggregate count, and the influenza type counts are converted into feature columns. One-hot encoding is used to generate dummy features for each county.

The weather data comes in the format of daily reports. This data is merged to match the weeks in the Influenza dataset and the average of the merged days is used as the value of the new data. The missing values in this dataset are replaced with the average value for that county for that week of the year ie a missing value in week 48 is replaced with the average of the values in the other week 48’s for that county in the dataset. Finally, because the weather of the week being forecast will not be known, weather features are added representing the average value for that week across the other years in the dataset.

The Google Trend data must be downloaded in periods spanning no more than 5 years. This causes the popularity rank data to normalize relative to the highest search popularity in that period, with that entry representing 100. These years have to be normalized back to being relative to the highest popularity period over the full 2009-2019 period being analyzed. In addition, Google Trends data is reported at the metropolitan area level instead of county. Because each metropolitan area includes multiple counties, an analysis of these areas is completed to sort each county into a metro area and then a county is assigned to each entry in the Google Trends dataset.

**Modeling and Evaluation**

**Deployment**

The model predicts the near-term laboratory confirmed cases of influenza to a high degree and is marked improvement over the base model. This model could be used to provide forecasts in the weekly CDC reports or used by local and federal government in deciding whether to activate special interventions such as increased public engagement about vaccination or accelerating school-based vaccination campaigns. Hospitals can use these forecasts to prepare themselves for any surges by ensuring appropriate staffing and stocking necessary supplies such as IV fluids and flu testing assays. Much of the data is gathered from surveys and datasets collected by the various federal government agencies and a data pipeline should be constructed allowing this data to come to the CDC directly instead of manually gathered. It must be noted that the model must be retrained frequently, as having the forecast can be correct but ultimately be too high because the interventions it triggered reduced the actual flu infection count. Frequent AB testing and other further research would be needed to tune for this effect.

*Ethical Considerations*

The model uses entirely anonymized aggregated data, so there is no confidentiality. However, demographic values such as racial composition were not included in the model because they are not available in sufficiently granular form. This means that the forecast model could unintentionally result in disparate impact on specific groups. In addition, the forecast model must be extremely sensitive to large deviations as any overallocation of resources of flu prevention would mean other clinical areas are being put on the sideline, while under-allocation due to reaction to a too-low forecast would result in unnecessary deaths associated with the flu.

1. <https://www.cdc.gov/flu/weekly/fluactivitysurv.htm> [↑](#endnote-ref-1)
2. <https://healthdata.gov/dataset/influenza-laboratory-confirmed-cases-county-beginning-2009-10-season> [↑](#endnote-ref-2)
3. <http://sitn.hms.harvard.edu/flash/2014/the-reason-for-the-season-why-flu-strikes-in-winter/> [↑](#endnote-ref-3)
4. <https://trends.google.com/trends/explore?geo=US-NY&q=flu> [↑](#endnote-ref-4)
5. <https://www.cdc.gov/flu/fluvaxview/1819season.htm> [↑](#endnote-ref-5)
6. <https://www.bea.gov/data/gdp/gdp-county> [↑](#endnote-ref-6)
7. <https://www.ers.usda.gov/data-products/atlas-of-rural-and-small-town-america/download-the-data/> [↑](#endnote-ref-7)