Predicting Pharmaceutical Demands to Treat Influenza at a Regional Hospital

Each fall, Western Slope Community Hospital (WSCH), a fictional hospital in Colorado, faces a sharp increase in the number of patients admitted with a common influenza virus. While hospital administrators recognize that flu cases increase seasonally, they do not have a systematic way to anticipate the number of cases day by day or week by week. Accurately predicting flu cases is essential because having sufficient antiviral medication on hand is critical for patient health. As Nicoll et al. (2012) and Dumiak (2012) wrote, effective antiviral treatment strategies are an indispensable part of addressing influenza cases. Nicoll et al. (2012) discussed the importance of preparing for large-scale viral infections, noting among essential steps local preparedness and antiviral treatment and vaccination strategies. Concerning the H1N1 influenza pandemic, Dumiak paraphrased Dr. Nahoko Shindo from the World Health Organization's (WHO) Global Influenza Programme, writing "available data show that antivirals helped to save lives" (2012, p. 800).

The WSCH CIO tasked the WSCH data analytics team with addressing the data problem of needing to accurately predict the demand for antiviral medication a week or two ahead of time so that the hospital would be adequately prepared for sudden increases in influenza cases. The team confirmed with the pharmaceutical department that having one week lead time would be sufficient for ordering antiviral pharmaceuticals in time to meet patient needs.

Literature Review

Several studies concerning predictive analysis for influenza outbreaks referenced Google Flu Trends (GFT). Pervaiz, Pervaiz, Rehman, and Saif (2012) explained that "the Google Flu Trends service was launched in 2008 to track changes in the volume of online search queries related to flu-like symptoms." Pervaiz et al. (2012) wrote that GFT was designed to indicate changes in the number of disease cases - but not as a system to detect epidemics. Even so, researchers found that GFT was useful in predicting influenza outbreaks. Malik, Gumel, Thompson, Strome, and Mahmud (2011) used data from the 2009 H1N1 pandemic waves in Manitoba to plot weekly counts of laboratory-confirmed H1N1 infections against three indicators: GFT data and two Emergency Department (ED) data points: the weekly count and the percentage of all ED visits treated as influenza-like illness (ILI) cases. They fitted a linear regression model separately for each indicator and found that all three indicators peaked one to two weeks earlier than the epidemic curve of cases confirmed by laboratories. Their best-fitting model for GFT data was ahead of the epidemic curve by two weeks, while their best-fitting models for each of the ED indicators was ahead by one to two weeks.

Dugas, Jalalpour, Gel, Levin, Torcaso, Igusa, and Rothman, (2013) similarly created multiple models to forecast United States influenza cases from 2004–2011. They wrote that their goal was "to provide individual medical centers with advanced warning of the expected number of influenza cases, thus allowing for sufficient time to implement interventions" (Dugas et al., 2013, p. 1). They used the weekly counts of laboratory-confirmed influenza cases, GFT data, meteorological data, and temporal information when creating their models. They trialed a few algorithms, including classical Box-Jenkins, generalized linear models (GLM), and generalized linear autoregressive moving average (GARMA). Dugas et al. (2013) found that a GARMA(3,0) forecast model with Negative Binomial distribution using ED and GFT data provided the most accurate influenza case predictions. They found that meteorological and temporal data did not improve predictions. They concluded, "integer-valued autoregression of influenza cases provides a strong base forecast model, which is enhanced by the addition of Google Flu Trends" (Dugas et al., 2013, p. 1).

Unfortunately, over time, GFT lost its ability to predict flu outbreaks. March 27, 2014, Arthur (2014) wrote that GFT had overestimated the number of flu cases for 100 of the previous 108 weeks. Leber (2014) similarly reported GFT's inaccurate predictions, noting that they were often no better than the Center for Disease Control's (CDC's) predictions. Martin, Xu, and Yasui (2014) wrote that GFT was re-calibrated in 2009 (after missing the first wave of the H1N1 pandemic in the United States) and prior to the 2013–2014 flu season (after overestimating the 2012–2013 flu season and predicting its peak three weeks late). Martin et al. (2014) suggested that the GFT modeling had weakened in part due to changes in users' search behavior and in part due to changes to the Google search algorithm. GFT data is no longer published and so was unavailable to the WSCH data analytics team. The team found that the CDC publishes weekly surveillance data (summarized nationally and regionally) representing instances of both lab-confirmed flu cases and ILI cases.

Analysis

Analyzing Data Sources

The WSCH data analytics team analyzed data available from the CDC's "National, Regional, and State Level Outpatient Illness and Viral Surveillance" page. The team used data representing Region 8 (Colorado's region) and identified a subset of variables that were appropriate for the predictive analytics project. The team noted that CDC data were reported in year-week intervals (such as 2019, week 3), a time frame that was appropriate for the current project.

The WSCH data analytics team downloaded all available data for Region 8 from the CDC's "National, Regional, and State Level Outpatient Illness and Viral Surveillance" page. The data were delivered as a zip compression file containing four Comma Separated Values (CSV) files with ILI data and confirmed flu cases data from 1997 to present. ILI data existed in a single file. It was reported as the percentage of physician visits related to ILI. Confirmed flu cases were represented in three CSV files: (1) data from clinical laboratories since the 2015 flu season, (2) data from public health laboratories since the 2015 flu season, and (3) combined data from both clinical and public health laboratories prior to the 2015 flu season. The data began in 1997. The data analytics team chose to use data from the last 10 flu seasons since the EMR system from which additional data would be extracted came online in early 2008.

The team extracted a subset of data from the CDC's Region 8 ILI data: the weekly summary information representing (1) weighted percentage ILI, (2) ILI total count, and (3) total patient count. The team also extracted a subset of data from the CDC's Region 8 lab-identified flu cases. The weekly summary information that was retained represented (1) the specimen count, (2) the count that tested positive for A, (3) the count that tested positive for B, (4) the percent positive, (5) the percent positive for A, and (6) the percent positive for B. These fields were computed for weeks from the 2015-2016 flu season on, since the clinical and public health data needed to be combined. They were also computed for earlier flu seasons because numbers were reported in finer detail (listed by flu subtypes). Finally, the team derived four weekly summary data points from WSCH's EMR system: (1) the number of ED cases, (2) the number of ED cases with ILI, (3) the Percent of ED cases that presented with ILI, and (4) the number of patients that began flu-specific antiviral medication (the target variable). The team considered a visit to be an ILI case if the patient's chief complaint included one of a handful of flu symptoms, including weakness, shortness of breath, cough, fever, and sore throat. See Table 1 for the complete list of variables included in the study.

Table 1

Variables Used for the Influenza Predictive Modeling project

Variable	Source	Raw or Derived?
Year	-	-
Week	-	-
EMR_Count_Total_ED_Cases	WSCH EMR	Derived: Weekly sum
EMR_Count_ED_Cases_with_ILI	WSCH EMR	Derived: Weekly sum
EMR_Pct_ED_Cases_with_ILI	WSCH EMR	Derived: Weekly ratio
EMR_Count_Patients_Start_Meds*	WSCH EMR	Derived: Weekly sum
ILI_Weighted_Percent	CDC ILI data	Raw
ILI_Count_ILI_Cases	CDC ILI data	Raw
ILI_Count_Total_Cases	CDC ILI data	Raw
LAB_Count_Total_Specimen	CDC Lab data	Derived**
LAB_Count_A_Positive	CDC Lab data	Derived**
LAB_Count_B_Positive	CDC Lab data	Derived**
LAB_Pct_All_Positive	CDC Lab data	Derived**
LAB_Pct_A_Positive	CDC Lab data	Derived**
LAB_Pct_B_Positive	CDC Lab data	Derived**

* EMR_Count_Patients_Start_Meds (the count of patients beginning antiviral medications) is the target variable.
** Varies by data source: Raw in CDC Clinical data since 2015; derived from CDC Public Health and Combined data. The raw CDC Clinical data and derived CDC Public Health data was combined together for data since the 2015 flu season,

The WSCH data analytics team created a composite data source with one row for each year-week. Columns represented the EMR's ED data (Total Count of ED cases, Count of ED cases with ILI, Percent of ED cases with ILI, and Count of Patients Beginning Antiviral Meds: the target variable), the CDC's ILI data (ILI Weighted Percent, ILI Patient Count, and ILI Total Patient Count), and the CDC's laboratory data (Lab Specimen Count, Lab A Positive Count, Lab B Positive Count, Lab All Percent Positive, Lab A Percent Positive, and Lab B Percent Positive). The Count of Patients Beginning Antiviral Meds was identified as the target variable (rather than a related measure such as doses of antiviral medications dispensed) so that each case treated as influenza would be counted only once.

Identifying Appropriate Predictive Models

As Mehler wrote (2017), specific predictive analytic problems require different algorithms. For example, Mehler suggested that classification algorithms are useful for questions concerning customer retention and recommendation systems, clustering algorithms are useful for segmentation, and regression algorithms are useful for predicting calendar-driven outcomes. Ray (2015) further clarified that regression algorithms are useful for forecasting and for discovering causal relationships between variables. In the current scenario, regression algorithms were determined to be the most useful choice.

Based on the accurate predictions Malik et al. generated with a linear regression model (2011), the team included a linear regression model in their project. Based on the work by Dugas et al. (2013), the team generated models using classical Box-Jenkins and generalized linear autoregressive moving average (GARMA) algorithms. In addition, the team noted that input variables were likely to be correlated and so they considered regression models that Ray (2015) stated were robust to multicollinearity. Following his guidance, they chose to build models using Ridge Regression and Lasso Regression. The data analytics team also decided to use a SAS Enterprise Miner Ensemble node to generate a model using the two models that performed the best.

Afzali, Gray, and Karnon (2013) emphasized the importance of validating and comparing models, writing "for the model to be a practical means of informing policy decisions, decision makers must have confidence that the model presents an accurate reflection… Central to these guidelines is a framework to improve the accuracy, and hence the credibility, of decision analytic models." (p. 86). The data analytics team decided to use SAS Enterprise Miner to generate and compare regression models in order to have confidence that the models used were more effective than their counterparts. The team planned to compare models of the same type in order to identify the most accurate model within a given type (such as Lasso regression models), as well as to compare models of different types to identify the most accurate predictive model overall.

Methodology

Data Acquisition

The WSCH data analytics team downloaded the CDC's weekly influenza report from the CDC's "National, Regional, and State Level Outpatient Illness and Viral Surveillance" page, pulling ILI and lab data for Region 8 from 2008 to present. It also exported weekly summary data from the EMR system from 2008 to present. Using Python, the team created a composite table that included columns of data from each of the sources (see Table 1 for the columns in the composite table). To have updated data for future predictions, the team also created an automated process to retrieve the previous week's data from the CDC's website and the WSCH EMR system. Data values were then appended to the data source table.

Configuring SAS Enterprise Miner and Generating Models

The WSCH data analytics team chose to use SAS Enterprise Miner for the influenza prediction project because SAS Enterprise Miner has several features that were important for the project. First, SAS Enterprise Miner easily connects with the data. Next, it has some nodes that use the predictive modeling algorithms that the team planned to use, including linear and Lasso regression models. As well, SAS Enterprise Miner has open source integration nodes that allow the team to include additional models written in R. As well, SAS Enterprise Miner has an Ensemble node that generates a predictive model based on existing models. Finally, it has a Compare Models node that makes it easy to compare predictive models' performances.

The WSCH data analytics team created a new project in SAS Enterprise Miner. Next, they created a data source object linked to the composite table containing relevant weekly summary CDC and EMR data. The data source was added to a new diagram and the variable representing the number of patients who began an antiviral medication was identified as the target variable. Output from the data source was linked to a StatExplore node to produce summary statistics for the variables. The variables were reviewed, and all were confirmed to have been identified as numeric. Variables with missing values were noted for later imputation. Brown (2016) wrote that skewness above 1 or below -1 represents highly skewed data. Variables with "highly skewed" data by that definition were also identified. A Graph Explore node was linked to the output of the StatExplore node to inspect the dataset; it provided a detailed view of the data. Data were inspected for erroneous data. Data that was suspected to be erroneous was updated (if possible) or imputed (if necessary) if confirmed to be erroneous.

The data were then partitioned into 70% training and 30% validation datasets using a Data Partition node. Since Regression nodes reject observations with missing values, and since they work best with normalized data, data were then imputed and transformed as needed before being used by the predictive models. Values were imputed with an Impute node configured to impute missing values using the mean (as all variables contained numeric values). Next, values were transformed. A Transform Variables node was added to the diagram, linked to the Impute node's output. In the Properties Panel, under "Train" properties, "Formulas" was selected to see histograms for variables with skewness greater than 1 or less than -1. Variables were transformed using transformations that were appropriate to reduce skewness. A second StatExplore node was added to the diagram receiving the output of the Transform Variables node to confirm the skewness was adequate. This second StatExplore node was also used to confirm that no variables had missing values after imputing occurred.

The output from the second StatExplore node was linked to a Control Point node, which linked to several models: linear regression (using a Regression node), classical Box-Jenkins (using an Open Source Integration node), GARMA (using an Open Source Integration node), Ridge (using an Open Source Integration node), and Lasso Regression (using an HP Regression node configured with the "Lasso" method).

Each of the models was run with a variety of configuration settings. Where appropriate, a SAS Code node was used to automate using different configuration settings. For each model type, the team used a Model Comparison node to compare different versions of the same type of model in order to identify the configuration settings that resulted in the best performances. For each model type, the configuration settings of the best performing model were used in the final SAS Enterprise Miner diagram. All of the finalized models were compared by linking them to a Control Point node, which was then linked to a Model Comparison node to compare the best version of each algorithm in order to identify the most accurate predictive model overall. The two top performing models were then linked to an Ensemble node, which was also linked to the Control Point node that linked to the Model Comparison node so that the Ensemble model could be compared with the other five models (see Figure 1).

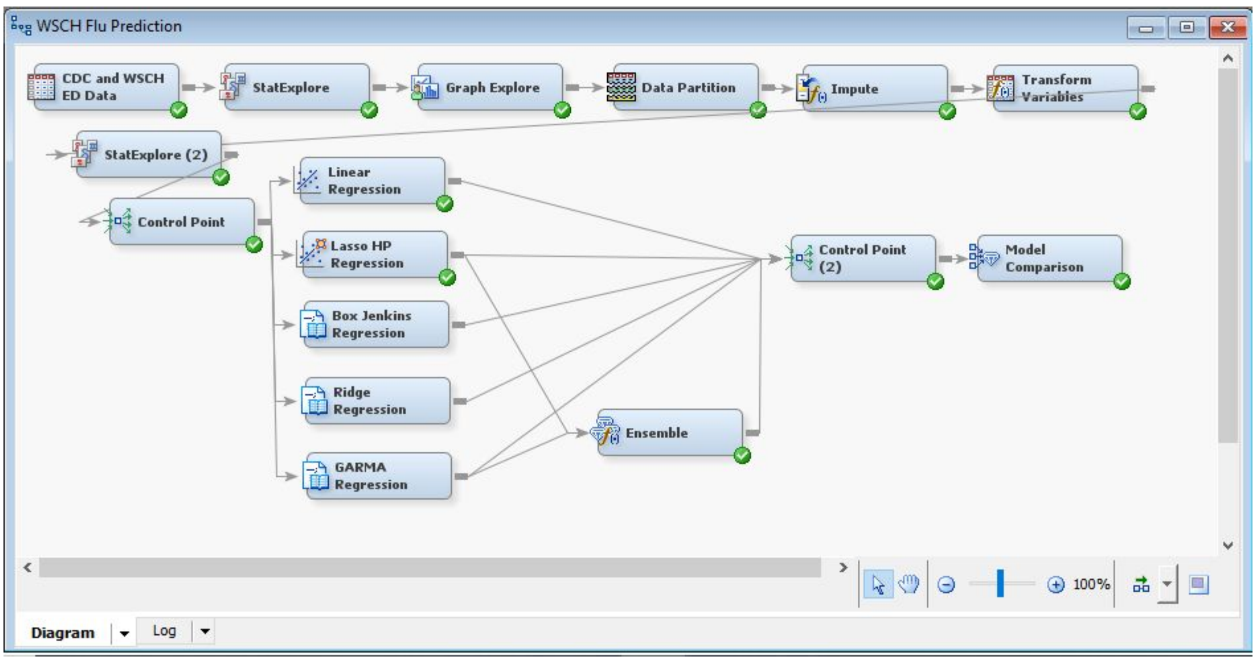

Figure 1. The final SAS® Enterprise Miner™ diagram, with data pre-processing, several models, and a model comparison node.

Conclusion

This predictive analytics project addressed the need for anticipating peaks in influenza cases ahead of time so that a hospital could be prepared to meet patient needs. Cases of influenza are notoriously difficult to predict. As the CDC wrote on its “Frequently Asked Flu Questions

2018-2019 Influenza Season" web page, "it is not possible to predict what this flu season will be like. While flu spreads every year, the timing, severity, and length of the season varies from one season to another."

While WSCH administration was most concerned about having an adequate supply of antiviral medication on-hand for influenza outbreaks, research to solve the problem identified that best-practice planning for outbreaks would require additional steps. Dugas et al. explained that an influenza forecast model "could increase planning capabilities beyond simply the next 24 hours, giving hospitals the crucial time needed to prepare for increased patient volumes whether through distribution or purchase of supplies, increased staffing, or opening additional annex areas to increase bed capacity" (2013, p. 3).

WSCH administration could use the predictive model in a variety of ways. Since the WSCG administration has an online reporting system for planning purposes, the number of anticipated flu cases could be added to the interface. Data from the hospital pharmacy's inventory system could also show the number of doses of antiviral medications on-hand, along with a note about the number of doses that are required for a given patient's treatment. The number of anticipated flu cases could also be integrated into a staff scheduling page, perhaps with an indicator if the number exceeded a threshold value, so that people involved in scheduling staff would know that additional staff would likely be needed. This indicator could also be useful if additional areas of the hospital needed to be made available to accommodate the increased patient load.

References

Afzali, H. H. A., Gray, J., & Karnon, J. (2013). Model performance evaluation (validation and calibration) in model-based studies of therapeutic interventions for cardiovascular diseases: A review and suggested reporting framework. *Applied Health Economics and Health Policy, 11*(2), 85-93.

Arthur, C. (2014, March 27). Google Flu Trends is no longer good at predicting flu, scientists find. *The Guardian*. Retrieved from https://www.theguardian.com/technology/2014/mar/27/google-flu-trends-predicting-flu

Brown, S. (2016, May 22). Measures of shape: skewness and kurtosis. Retrieved from https://brownmath.com/stat/shape.htm

Center for Disease Control and Prevention. (2017, October). CDC FluView weekly report: National, regional, and state/jurisdiction level outpatient illness and viral surveillance application quick reference guide. Retrieved from https://gis.cdc.gov/grasp/fluview/FluViewPhase2QuickReferenceGuide.pdf

Center for Disease Control and Prevention. (n.d.). Flu activity & surveillance. Retrieved from https://www.cdc.gov/flu/weekly/fluactivitysurv.htm

Center for Disease Control and Prevention. (n.d.). Frequently asked flu questions 2018-2019 influenza season. Retrieved from https://www.cdc.gov/flu/season/flu-season-2018-2019.htm

Center for Disease Control and Prevention. (n.d.). National, regional, and state level outpatient illness and viral surveillance. Retrieved from https://gis.cdc.gov/grasp/fluview/fluportaldashboard.html

Center for Disease Control and Prevention. (n.d.). Weekly U.S. influenza surveillance report. Retrieved from https://www.cdc.gov/flu/weekly/

Chisholm, M. (2015, July 14). 7 phases of a data life cycle. *Information Management*. Retrieved from https://www.bloomberg.com/professional/blog/7-phases-of-a-data-life-cycle/

Dumiak, M. (2012). Push needed for pandemic planning. World Health Organization. *Bulletin of the World Health Organization, 90*(11), 800-1. doi:http://dx.doi.org.csuglobal.idm.oclc.org/10.2471/BLT.12.021112

Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., and Rothman, R. E. (2013). Influenza Forecasting with Google Flu Trends. *PLoS ONE 8.2*: E56176. Web.

Leber, J. (2014, March 13). The failures Of Google Flu Trends show what's wrong with big data. Retrieved from https://www.fastcompany.com/3027585/the-failures-of-google-flu-trends-show-whats-wrong-with-big-data

Malik, M. T., Gumel, A., Thompson, L. H., Strome, T., & Mahmud, S. M. (2011). Google Flu Trends and emergency department triage data predicted the 2009 pandemic H1N1 waves in Manitoba. *Canadian Journal of Public Health, 102*(4), 294-7. Retrieved from https://csuglobal.idm.oclc.org/login?url=https://search-proquest-com.csuglobal.idm.oclc.org/docview/884329236?accountid=38569

Martin, L. J., Xu, B., & Yasui, Y. (2014, December 31). "Improving Google Flu Trends Estimates for the United States through Transformation." *PLoS ONE* 9.12 (2014): E109209. Web.

Mehler, S. M. (2017, April 3). The ultimate guide for choosing algorithms for predictive modeling. Retrieved from https://www.analyticbridge.datasciencecentral.com/profiles/blogs/the-ultimate-guide-for-choosing-algorithms-for-predictive

Nicoll, A., Brown, C., Karcher, F., Penttinen, P., Hegermann-Lindencrone, M., Villanueva, S., . . . Nguyen-Van-Tam, J.,S. (2012). Developing pandemic preparedness in Europe in the 21st century: Experience, evolution and next steps. *Bulletin of the World Health Organization, 90*(4), 311-317. Retrieved from https://csuglobal.idm.oclc.org/login?url=https://search-proquest-com.csuglobal.idm.oclc.org/docview/1284079739?accountid=38569

Pervaiz, F., Pervaiz, M., Rehman, N. A., & Saif, U. (2012). FluBreaks: Early Epidemic Detection from Google Flu Trends. *Journal of Medical Internet Research, 14*(5), 19. https://doi-org.csuglobal.idm.oclc.org/10.2196/jmir.2102

Ray, S. (2015, August 14). 7 types of regression techniques you should know! Retrieved from https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/

Sarma, K. (2017). *Predictive modeling with SAS Enterprise Miner: Practical solutions for business applications* (3rd ed.). Cary, North Carolina: SAS Institute

SAS Institute Inc. SAS Enterprise Miner. [Computer software]. (2014). Retrieved from https://odamid.oda.sas.com/SASODAControlCenter/

SAS Institute (2017). Advanced predictive modeling using SAS Enterprise Miner. [Course Notes] ISBN 978-1–63526-115-8