Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

Susannah Paul	**04/21/2017**
[Student's name typed]	Date

Identifying Change Points and Forecasting Influenza Trends Using Diverse Influenza-like Illness Surveillance Data Capture Mechanisms in the City of Houston, Texas (2012-2016)

By

Susannah Paul

Master of Public Health

Epidemiology

Roberd M. Bostick, MD, MPH

Committee Chair

Eunice Santos, MPH

Committee Member

Identifying Change Points and Forecasting Influenza Trends Using Diverse Influenza-like Illness Surveillance Data Capture Mechanisms in the City of Houston, Texas (2012-2016)

By

Susannah Paul

B.S., Biology and Business Administration

The University of Texas at Dallas

2013

Thesis Committee Chair: Roberd M. Bostick, MD, MPH

An abstract of

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

In partial fulfillment of the requirements for the degree of

Master of Public Health

in Epidemiology

2017

Abstract

Identifying Change Points and Forecasting Influenza Trends Using Diverse Influenza-like Illness Surveillance Data Capture Mechanisms in the City of Houston, Texas (2012-2016)

By Susannah Paul

Background: Because influenza activity in a region is influenced by multiple factors, such as vaccine effectiveness, virus mutations, and travel, it is difficult to anticipate or identify significant activity increases early. Reports from traditional surveillance systems, which report data from emergency centers or providers, are accurate but delayed because of delays in patients seeing a provider or waiting for laboratory confirmation. Using novel surveillance methods for complementary information and combining available historical data can lead to earlier detection of influenza activity increases and decreases. Anticipating a surge would give public health professionals more time to prepare for a rise in cases and increase prevention efforts to reduce the risk of an epidemic.
Methods: Our objective was to investigate influenza activity in the City of Houston, by analyzing influenza-like illness data from diverse data capture mechanisms, from week 27 of 2012 through week 26 of 2016. Change point analysis was used to identify significant increase and decrease change points within each data source. ARIMA models were fitted for each source and used to estimate forecasts for influenza-like activity for the subsequent 10 weeks.
Results: All sources except for Flu Near You contained at least one significant increase and one significant decrease within each time interval. Overall, Athena, ILINet, and ER Centers resulted in similar start and end dates of the influenza season. Multiple, gradual changes within the typical influenza season and during non-seasonal time were identified. Forecasted estimates had wide confidence intervals with lower bounds below zero. The forecasted trend direction differed by data source, resulting in lack of consensus about future influenza activity.
Conclusion: The similarities in trend and timing of identified change points from outpatient information, ER Centers patient data, and ILINet influenza-like illness provider surveillance supports potential use of diverse data capture mechanisms to enhance influenza surveillance. Pooling the significant change points results in a comprehensive trend pattern identification for influenza activity. Though the forecasted estimates did not agree on trend direction and would be inaccurate for long-term predictions, pooling the predictions could be helpful in the short term if more historical information and influential variables are considered.

Identifying Change Points and Forecasting Influenza Trends Using Diverse Influenza-like Illness Surveillance Data Capture Mechanisms in the City of Houston, Texas (2012-2016)

By

Susannah Paul

B.S., Biology and Business Administration

The University of Texas at Dallas

2013

Thesis Committee Chair: Dr. Roberd Bostick, MD, MPH

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Public Health

in Epidemiology

2017

Acknowledgements

I acknowledge the Office of Surveillance and Public Health Preparedness as well as the Bureau of Laboratory Services at the Houston Health Department. This thesis uses data from Flu Near You, an application where voluntary users from the public can self-report symptoms weekly, that resulted from the collaboration of epidemiologists at Harvard, Boston Children's Hospital, and The Skoll Global Threats Fund. I also appreciate the Texas Department of State Health Services for access to the ILINet data.

Thank you to Eunice Santos, epidemiologist for Houston Health Department for the weekly conference calls and guidance. I also thank Roberd M. Bostick, MD, MPH for serving as thesis faculty advisor.

Finally, I thank my family for the continuous encouragement and patience throughout this process.

Table of Contents

Introduction

Background and the Burden of Influenza

Syndromic surveillance is used to monitor and evaluate trends for numerous illnesses and disease-causing agents. Initially, syndromic surveillance systems were created to assess information on the size, spread, and trends of cases for early detection of outbreaks due to biological terrorism agents (Henning, 2004). This was not proposed to replace direct physician reporting of cases of public health importance but intended as a complementary tool. In the 2000 strategic plan, the Centers for Disease Control and Prevention (CDC) called for new mechanisms for detecting suspicious events (Khan & Levitt, 2000). The 2001 terrorist attacks and anthrax outbreak from the release of *Bacillus anthracis* spores in the United States (U.S.) spurred the public health community to implement syndromic surveillance systems across the U.S. (Henning, 2004). Although the initial purpose was detection of large-scale bioterrorism agents, syndromic surveillance has expanded quickly to monitor other disease-causing agents due to its capability and flexibility in identifying disease clusters before diagnosis and lab confirmations are reported. Technological advancements such as the use of electronic health records have prompted systems to adapt syndromic surveillance methods to detect influenza and influenza-like illness activity.

The burden of influenza in the U.S. varies widely based on the circulating strains, vaccine efficacy, vaccine effectiveness, and other factors. Seasonal influenza infects 5-20% of the U.S. population annually while over 200,000 people are hospitalized with influenza-related complications (Davidson, Haim, & Radin, 2015). Young children, pregnant women, and older adults are considered particularly vulnerable populations.

From 2003 to 2012, an average of 6,514 infants <12 months were hospitalized in the U.S. due to influenza or illnesses complicated by influenza, with 75% of hospitalizations occurring in healthy infants. The proportion of infants with high risk conditions admitted to the intensive care unit (ICU) or with respiratory failure was 2-3 times higher than compared to otherwise healthy infants. Infants less than 6 months were 40% more likely to be admitted to the ICU (Chaves et al., 2014). Influenza has a high impact on infants, and those under 6 months are not recommended for influenza vaccination. This bolsters the need for additional syndromic influenza surveillance. Syndromic surveillance data helps the public health community prepare for increases in influenza activity, especially among populations with low influenza vaccination rates.

In 2009, the H1N1 strain began an influenza pandemic (Caspard et al., 2016). The strain that caused the 2009 pandemic is now considered a seasonally circulating strain. The cases initially discovered via lab confirmations in April 2009 resulted in a declaration of Public Health Emergency of International Concern by the World Health Organization (WHO). There were 18,500 laboratory-confirmed reported deaths caused by this strain between April 2009 and August 2010 with the true mortality count likely higher (Dawood et al., 2012). As the pandemic spread, schools closed and people enacted social distancing measures (CDC, 2010). This pandemic physically limited many people, but it also had heavy socioeconomic impacts compared to other influenza seasons. For example, in the Republic of Korea, total socioeconomic burden was estimated to be $1,581.3 million (10-90%: $1,436.0 – $1,808.3) compared to $42.3 million (10-90%: $31.5 – $53.8) in the 2008-2009 season. Longer sick leaves and high diagnostic costs contributed the most to the increase (Suh et al., 2013). Public health professionals have

been continuously monitoring and developing influenza surveillance systems to prevent another pandemic of this extent.

Although influenza peaks typically occur between December and February, because of the changes in strains each year, influenza peaks can occur earlier or later and with higher magnitudes than expected resulting in epidemic levels. For example, during the influenza season of 2013, there were spot vaccine shortages. In the United States, the deaths of 29 children and 5, 249 hospitalizations by the second week in January made obvious the increasing intensity of the influenza season. This increasing case count led people to rush to receive influenza vaccines leading to shortages and difficulties in obtaining flu vaccines in certain regions (Phyllis, 2013). Anticipating a surge in influenza using data on influenza-like illness gives health care professionals more time to prepare for a rise in cases and increase prevention efforts to reduce the risk of an influenza epidemic.

Influenza-like Illness and Syndromic Surveillance Systems

Influenza-like illness (ILI) is defined in the U.S. as fever with a temperature of 100℉/37.8℃ or greater and a cough or sore throat without another known cause (Centers for Disease Control and Prevention, 2016). Information on ILI has been used as an indicator of influenza activity. This information is not only useful to monitor influenza activity, but also to forecast future waves of influenza as well (Ong, et al., 2010). ILI data are usually gathered from sentinel providers or health care facilities utilizing electronic health records. However, utilization of innovative data capture mechanisms and platforms such as social media or self-reporting smartphone applications can enrich the ILI data collected and make influenza surveillance and preparedness more efficient.

The consistent flow of clinical data from emergency departments (ED) is standard for an accurate estimation of influenza activity in a specific region. Two large surveillance networks, the North American DiSTRIBuTE network and the European Triple S system, collect large-scale emergency department-based influenza and ILI syndromic data. Clinical data is usually collected based on chief complaint or diagnosis (Hiller, Stoneking, Min, & Rhodes, 2013). Emergency medical service (EMS) syndromic influenza data can be valuable to capture more moderate to severe cases of influenza. Researchers in Europe compared timeliness and validity of information collected from an Austrian Emergency Medical Dispatch Service, Austrian and Belgian ambulance services, and a Belgian and Spanish emergency department during the 2009 influenza pandemic. Results indicated that the emergency department data displayed the most favorable validity and timeliness (Rosenkötter et al., 2013).

Influenza-like illness data from outpatient data provided by health care facilities is also useful in estimating influenza trends. A study from October 2009 to July 2010 with 38 outpatient practices showed that observed trends were consistent with already established syndromic and laboratory systems. This shows the feasibility of using outpatient information to estimate influenza in a community (Fowlkes et al., 2013).

The use of big data is a relatively novel approach to influenza surveillance. Big data is a general term for enormous data sets with large, varied, and complex structures (Sagiroglu & Sinanc, 2013). Internet search queries are a potential source to show influenza activity trends in a region. Internet utilization is globally widespread and there are multiple platforms, such as cellular phones or computers, from which to access the internet. Search engines are a standard resource for retrieval and dissemination of

information. People may conduct internet searches prior to visiting a health care provider. Tracking these trends could capture earlier surges in activity. A study in South Korea compared ILI data from the Korea Centers for Disease Control and Prevention to ILI related queries on the local search engine, Daum. Researchers concluded that methods showed strong, positive correlation with national influenza surveillance data in South Korea (Seo et al., 2014).

Google launched Google Flu Trends (GFT) hoping that big data would effectively estimate influenza spread faster than traditional surveillance systems. Evaluations of GFT results showed some significant errors that reduced the tracking accuracy of laboratory-confirmed cases. GFT missed the first wave of the 2009 influenza A/H1N1 pandemic and overestimated the A/H3N2 epidemic in 2013. However, combining Google's data with traditional surveillance system data improves results. A test of an empirical model that combined Google's estimates with CDC data showed a 2.1% reduced error in case prediction compared to GFT alone (Davidson et al., 2015). GFT is also useful as complementary information when trying to forecast cases of influenza (Dugas et al., 2013). Passive search queries and actively collected crowd-sourced data contain a lot of noise that may reduce reliability, but the volume and velocity of big data collection can decrease the time from evaluation to action with continuous calibration (Hay, George, Moyes, & Brownstein, 2013). This demonstrates the potential usefulness of big data when combined with traditional clinical systems.

With advancements in technology and social media, there is an effort to utilize new resources to track health-seeking behaviors associated with influenza. The goal is to decrease the time between preparation for an influenza surge and the actual wave of

influenza activity itself. With enough preparation, it may be possible to reduce influenza-associated morbidity and mortality. Although traditionally ILI is defined with clinical symptoms, with the role of technology in monitoring health behaviors, measures such as influenza-related search queries can also behave similarly to ILI clinical symptoms. These measures in conjunction with traditional ILI data can also be used to monitor, analyze, and forecast influenza activity (Signorini, Segre, & Polgreen, 2011).

Gaps in Surveillance to Reduce Influenza Risk

Influenza vaccines are widely used to reduce the burden of seasonal influenza. People six months and older are encouraged to receive the vaccine annually. However, because of low vaccine uptake rates, influenza surveillance and preparedness also plays a vital role in reducing this burden. In a study of 16,683 patients with health care providers participating in surveillance to monitor acute respiratory illness, researchers found that from August 2010 to December 2013, only 30.4% of influenza test-negative and 18.6% of test-positive patients reported influenza vaccination (*Vaccine*, 2016). Overall, vaccine coverage for other diseases is high in the U.S. due to school and university entry requirements. For example, measles, mumps, and rubella (MMR) vaccine coverage ranged from 87.1% to 99.4% and hepatitis B (HepB) ranged from 90.3% to 99.6% depending on the state (National Center for Immunization and Respiratory Diseases, 2016). As of now, although some work environments require annual vaccination, there are no statutes that require people to get an influenza vaccine. Also unlike most other vaccinations, the contents of the influenza vaccine change every year as the circulating strains of the virus also adapt or change. Though the seasonal influenza vaccine exhibits varying efficacy and effectiveness depending on the season, it is an effective tool to

reduce the burden of season influenza compared to not receiving the vaccines. Regardless, many still do not receive the vaccine. It is unlikely that the influenza vaccine coverage rates will reach that of the more common child vaccinations such as MMR or hepatitis B vaccines. Therefore, influenza surveillance is valuable for preparedness in event of possible epidemics.

Even with high influenza vaccination coverage rates, there is still an immense need for influenza surveillance. The development and production of influenza vaccines results in an intelligently designed, educated guess. Researchers present three to four candidate vaccine viruses that they anticipate will be most likely to spread in the upcoming influenza season. Due to how quickly influenza viruses can mutate or adapt, it is possible that protection against another unanticipated circulating strain is low. A study to evaluate quadrivalent live attenuated influenza vaccine (LAIV) effectiveness in children 2-17 years in the U.S showed that LAIV provided significant protection against "B/Yamagata influenza but not against A/H1N1pdm09" in 2013-2014. In contrast, the inactivated influenza vaccine was effective against both strains (Caspard et al., 2016). Based on these results, CDC rescinded their LAIV preference in 2015 (Smith et al., 2016). Influenza vaccine development is an iterative process and yet, even after approved release of a vaccine, the effectiveness may be much lower than in controlled environments. The entire process starts months before the anticipated influenza season so that manufacturers have time to produce the vaccines before outbreaks occur. As seen in previous years, it is possible that a vaccine can offer low protection for certain strains that were not anticipated to be prevalent or even due to mutations in vaccine strains (Skowronski et al., 2014). Therefore, even if there is high influenza vaccine coverage, it

is still possible to have high influenza activity or outbreaks. Researchers, manufacturers, distributers, and health care professionals can benefit from a quicker determination of influenza outbreaks so that they can anticipate if extra vaccines are needed. The vaccine manufacturing and distribution process is not instantaneous. Complementing traditional influenza surveillance systems with all available data capture mechanisms leads to earlier detection of influenza activity increases. This gives more time to prepared and respond to reduce the influenza burden. Influenza surveillance is still a public health necessity throughout all levels of vaccine coverage. Further studies can utilize multivariate methods that incorporate measures of vaccine effectiveness to explore how it affects influenza trends.

With consistent efforts towards improving influenza-related surveillance systems, the public health community has improved at anticipating influenza outbreaks quickly. However, awareness with traditional surveillance systems usually comes after enough people have become cases and exceeded the threshold. Time has passed between initial infection and lab confirmation of infection. Also, these systems only capture those who think they are sick enough to visit a provider and depend on enough providers participating and reporting in a timely manner. Typically, automated systems for syndromic surveillance currently do not identify change points with confidence intervals nor forecast or predict future influenza trends. Real-time, local data related to influenza makes this kind of analysis possible. In Hong Kong, authors reported development of systems capable of forecasting seasonal influenza epidemics in real-time in subtropical and tropical regions where epidemics can occur throughout the year. They retrospectively forecasted on non-seasonal influenza epidemics. Average forecast accuracies were 31%

with leads of one to three weeks (Yang, Cowling, Lau, & Shaman, 2015). Although not automated, peak predictions have been estimated in the past. For the 2003-2008 influenza seasons in New York City, researchers retrospectively forecasted weekly local influenza infection rates up to seven weeks in advance of the actual peak (Shaman & Karspeck, 2012).

Growing international connections and increased travel between countries increases the local risk of influenza. However, restricting air travel is less effective compared to local transmission reduction interventions (Viboud, Miller, Grenfell, Bjrnstad, & Simonsen, 2007). It is important to think about data sources innovatively so that as much information as possible can be used to alert the health care community of potential outbreaks locally first before the geographic area of the spread reaches too far. This reduces the response time so that health care professionals are better and earlier prepared. Preparedness or response efforts such as organizing adequate staffing, disseminating accurate information quickly, or prepping medical supplies or medications can benefit from timely alerts of increased local influenza activity. The speed at which an influenza increase is identified affects how quickly public health professionals can respond appropriately. It is important to utilize complementary systems with diverse data structures and sources to increase the efficiency of reporting, preparedness, analysis, identification, and response as it all relates to influenza.

Methods

Syndromic Surveillance Sources and Population

We obtained data from five different sources with diverse data capture mechanisms. Information included weekly aggregate ILI counts of the residents in the City of Houston jurisdiction. We analyzed data from July 2012 through June 2016 or week 27 of 2012 through week 26 of 2016. Analyzing diverse data capture mechanisms can give more information on the perceived severity of influenza in the population. For example, while emergency department data may represent mostly cases severe enough to warrant a visit to the hospital ER, outpatient data may capture milder cases.

GFT followed health behaviors online by tracking 40 different search queries related to influenza. This tool was used to predict prevalence from aggregated user data. Estimates for total number of search queries or information from the 2015-2016 season were not available and therefore, this source was not analyzed using change point analysis or forecasting.

Epidemiologists in partnership with Skoll Global Threats Fund, Boston Children's Hospital, and the American Public Health Association created Flu Near You. The tool is available as both a web-based platform and a phone application. Users can self-report various symptoms on a weekly basis. The number of responses with symptoms related to influenza-like illness are aggregated weekly over the total number of responses. Flu Near You opened to the public in November of 2011.

The Houston Health Department also receives data from the Athena Network, a group of healthcare facilities and providers. A billing company pulls patient records from outpatient clinics to produce weekly aggregates for ILI. When a patient visits an

outpatient clinic within the network in Houston, the international classification of diseases (ICD) codes are categorized as ILI based on the syndrome definition. This data is aggregated and recorded weekly and is not considered collected in real-time.

The Epidemiology and Prevention Branch in the Influenza Division at the CDC collects and compiles data on influenza-like illness in the United States. This is a collaborative effort between the CDC and health care providers and facilities. Health care providers report weekly on the total number of patients seen and the proportion of patients with influenza-like illness by age group. For comparability, ILINet will be used as the "gold standard" to which other sources are compared since it is one of the most comprehensive ILI surveillance systems in the U.S.

The Office of Public Health Preparedness and Surveillance within the Houston Health Department receives real-time data transferred daily from hospital emergency centers. When a patient is registered in an emergency center, each record is coded into a syndrome based on the chief complaint. ILI is coded and analyzed by someone using computer software to identify activity above the threshold. The threshold is calculated as the average ILI occurrence in the past three years from non-influenza season months (June-September).

The influenza season start, duration, and peak is determined each year by surveillance depending on timing of surges in activity. Historically, an influenza season is October through May because these months usually contain higher influenza activity. It is important to understand the influenza season period because preparedness and response activities are all coordinated around these time intervals of expected influenza activity surge.

The specific outcome for our analysis depends on type of surveillance mechanism. However, it is broadly defined as the percentage of ILI observations among the total observations in a week. Since GFT did not have the total number of search queries available, this source was not analyzed using change point analysis but counts are available for comparison.

The average ILI count and ILI percentage for October-May and June-September, peak week of highest ILI count and ILI percentage, and minimum and maximum ILI percentages are calculated and graphed to see if activity peaked around the same weeks among the diverse data capture mechanisms (**Table 1**, **Figure 1**). This initial analysis is conducted to visually compare trends and seasonal patterns. Analyses were completed using R 3.2 with R Studio, Change Point Analyzer 2.3, and Microsoft Excel.

Imputing Missing Values

There were 21 (10%) missing observations from ILINet. ER Centers data also showed one missing observation (0.5%). ILI percentages were not available for these weeks. The missing data is estimated using predictive mean matching. The *'mice'* package in R imputes missing values with plausible values using the predictive mean matching method. It uses an algorithm that pulls information from other values in the specified variable to predict possible values. Predictive mean matching estimates a linear regression for observed values. Then, it picks a value randomly from the posterior predictive distribution of the coefficients of the previous regression to produce a new set of coefficients. These coefficients predict values for all observations. For each missing observation, we pick a set of observations with predicted values close to predicted values

of the missing observations. We randomly choose one observed value to replace the missing observation (Allison, 2015).

CPA Methods: Cumulative Sum and Bootstrapping

Change point analysis is used to detect whether a change has taken place(Taylor, 2000). It can be used to detect multiple, subtle changes which is better suited for assessing increases or decreases both during and after the influenza season. With the methods used, each change detected is likely to be a real change.

Cumulative sum control charts are used to monitor the mean influenza-like activity of a given year based on weekly activity. This method can detect small changes because it utilizes current and previous week observations instead of averaging each week independently. CUSUM charts rely on an accurate target value and standard deviation. The cumulative sum for each week is the sum of the differences between the weekly ILI percentage and the average added to the previous weekly cumulative sum ($S_i = S_{i-1} + (X_i - \bar{X})$, where X_i = weekly percentage of ILI, S_{i-1} = previous week's cumulative sum, $\bar{X}$ = average). **Figure 3** contains CUSUM charts by time interval for the data sources. If there is no departure from the mean, the slope will be relatively flat. When the CUSUM slopes down, it indicates a period when percentages are below the mean. Upward slopes indicate a period when percentages are above the mean. Sudden changes in slope indicates a shift in the mean. We can visually note these points of interest but this is a subjective assessment (**Figure 3**). Bootstrapping to find confidence levels makes pinpointing actual changes more objective. The mean square error (MSE) is an estimator of when the change occurred. **Equation 1** contains the formula for calculating the MSE.

(Equation 1) MSE = $\sum_{i=1}^{m} (X_i - \overline{X_1})^2 + \sum_{i=m+1}^{n} (X_i - \overline{X}_2)^2$, where

$$\overline{X_1} = \frac{\sum_{i=1}^{m} X_i}{m} \text{ and}$$

$$\overline{X}_2 = \frac{\sum_{i=m+1}^{n} X_i}{n-m}.$$

After cumulative sum charting, bootstrap samples are generated to calculate a confidence level for each change. These samples represent random data samples that mimic the behavior of the cumulative sum if there has been no change. By performing 1000 iterations, we estimate how much the magnitude of change would vary if no change has occurred. Bootstrap samples are randomly reordered without replacement. The cumulative sum (S^0_n) and cumulative sum range (S^0_{diff}) are calculated for each of the samples. The percentage of times the $S_{diff} > S^0_{diff}$ is the confidence level. The predetermined threshold confidence level is 90%. A 95% confidence interval is used around the estimated change point week (Baddour, Tholmer, & Gavit, 2009).

ARIMA Modeling and Forecasting

Autoregressive integrated moving average (ARIMA) models are one of the most commonly used approaches to describe autocorrelations in the data and forecast time series (Hyndman & Athanasopoulos, 2014). The model requires time series data and autocorrelations that remain constant over time. The forecasting equation **(Equation 2)** is a univariate linear equation where the predictors are the lags of the variable and/or lags of the forecast errors and/or a possible constant (**Equation 2**). The lagged values of the variable are the autoregressive component while the lags of the forecast errors represent the moving average terms. The model is integrated if the time series must be differenced to be considered stationary.

Stationarity

ARIMA modeling requires stationarity. The mean and variance of a stationary time series do not depend on the time the series is observed. The ILI observations are not stationary because they have seasonality and possible trend. The data also have a white noise component, which is a stationary series because it looks the same at any time. One way to make a non-stationary time series stationary is by differencing. Differencing stabilizes the mean and removes the trend and seasonality components. It is the change between consecutive observations and so will result in one less observation since the difference cannot be calculate for the first observation. The time series observations used here only required first-order differencing and so second-order differencing will not be discussed. To make the decision whether to difference or not more objective, augmented Dickey-Fuller (ADF) tests were used to test if each of the data source time series variables were stationary or required differencing. The null hypothesis for this test is non-stationarity and any p-values above 0.05 indicate need for differencing. This test showed that the Athena data needed differencing. Another unit root test, the Kwiatkowski-Phillips-Schmidt-Shin (KPPS) test provided the same conclusion on stationarity for three out of four of the data sources. ILINet was found to be stationary with the ADF test and non-stationary with the KPSS test. Each of the sources were also tested using seasonal root tests to determine the appropriate number of seasonal differences required. None of the time series required seasonal differencing (**Table 2**).

Fitting an ARIMA Model

ARIMA models consist of parameter or coefficient estimates and determination of model orders, ARIMA (p,d,q)(P,D,Q)m. P,D,Q corresponds to the seasonal orders while

p,d,q represents the orders of the non-seasonal model. The order notation p is the autoregressive order, d, is the order of differencing, and q is the order of moving average. The m is equal to the number of periods per season.

The '*forecast'* package in R has some useful functions for determining the appropriate ARIMA models for each of our time series variables. The *auto.arima()* function estimates the parameters and model orders using maximum likelihood estimation (MLE). This method finds parameter values that maximize the probability of obtaining the observed data. Good ARIMA models also minimize the Akaike's information criteria (AIC) values. First, the number of differences, d, is determined with KPSS tests. After differencing d times, p and q are chosen by looking for the lowest AIC_c value. A constant, c, is included if d = 0 or d=1 if it improves the AIC value. The stepwise procedure varies the p and/or q by one and decides whether to include or exclude the constant. These steps are repeated until no lower AIC_c value is found. After the best model is found, ACF plots of the residuals and portmanteau Ljung-Box tests check the residuals. If the residuals appear to be white noise, this is the best model for forecasting.

Obtaining Point Forecasts from ARIMA Models

The general forecasting equation can be written as:

(Equation 2) $\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$ where,

θ = moving average parameters of order q,

ϕ = autoregressive parameters of order p,

$\hat{y}_t$ = prediction estimates at time t,

y_{t-p} = lagged values of y, and

e = error term.

The '*forecast'* package in R also provides a function called *forecast ()* that can be used once an appropriate model has been fitted. We obtain point estimates by replacing *t* with *t + h* in the equation, where *h=1,2….n* future forecasts. The future errors are replaced by zero and past errors are replaced by residuals. This function also provides forecast intervals at the 80% and 95% levels by default. The 95% forecast interval is calculated by $\hat{y}_{t+h|t} \pm 1.96 \sqrt{v}_{t+h|t}$, where $\hat{\sigma}$ is the standard deviation of the residuals and v is the estimated forecast variance.

Results

Identifying and Comparing Significant Change Points

Table 1. Peak Weeks per Year and ILI Percentages by Data Capture Mechanism					
Weeks	**Data Source**	**Peak Week[4]**	**Peak ILI %[4]**	**Mean[4]**	**Mean (Oct-May)**
2012-2013					
	Flu Near You	2012 45th	28.57	3.29	3.84[1]
	Athena	2012 50th	0.66	0.24	0.26[1]
	ER Centers	2012 51st	5.25	1.94	2.41[1]
	ILINet	51st 2012	2.62	0.72	1.01[3]
2013-2014					
	Flu Near You	2013 51st	10.29	2.22	2.48[1]
	Athena	2013 51st	0.48	0.15	0.17[1]
	ER Centers	2013 51st	6.59	1.87	2.35[1]
	ILINet	2013 51st	3.73	0.67	0.96[1]
2014-2015					
	Flu Near You	2014 37th	9.43	2.77	3.56[2]
	Athena	2014 50th	0.21	0.09	0.11[2]
	ER Centers	2014 51st	3.80	1.55	2.00[2]
	ILINet	2014 46th	1.48	0.28	0.41[2]
2015-2016					
	Flu Near You	2016 25th	6.9	1.90	2.12[3]
	Athena	2016 12	0.37	0.13	0.14[3]
	ER Centers	2016 9th	2.76	1.24	1.49[3]
	ILINet	51st 2015	0.96	0.18	0.23[5]

[1]Mean is calculated using weeks 40 through 22.
[2]Mean is calculated using weeks 40 through 21.
[3]Mean is calculated using weeks 39 through 21.
[4]The time interval is week 27 through week 26 of the following year.
[5]Mean is calculated using weeks 38 through 20.

Preliminary analysis shows the potential compatibility of the diverse data capture mechanisms and the times of peak influenza activity **(Figure 1, Figure 4, Table 1, Table 3)**. From analysis using change point analysis, at least two significant change points are identified per season per data capture mechanism. The exception is Flu Near You years 2012-2013, 2013-2014, and 2015-2016.

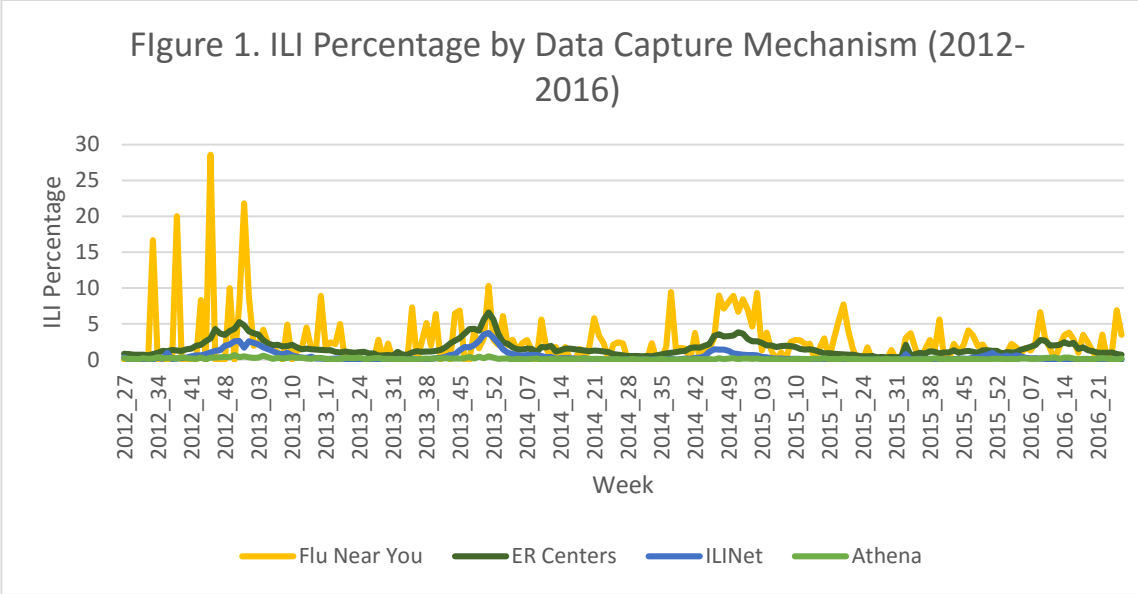


Figure 1. The above figure shows ILI percentage by data capture mechanism from week 27 of 2012 through week 26 of 2016. Flu Near You shows the greatest variability and highest ILI percentages. Athena outpatient data has the lowest percentages. Athena, hospital ER Centers, and ILINet show four clear seasonal peaks. Peaks are visible for all sources at similar times in the season. Any outliers or missing observations were removed from the data.

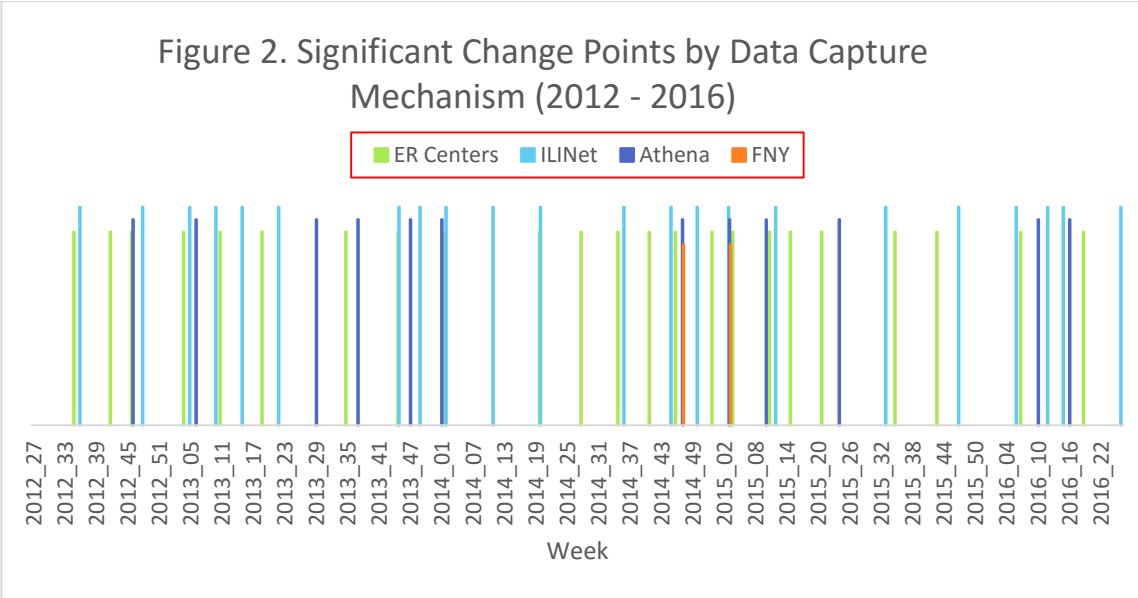


Figure 4. Each line represents an estimation of when a significant change occurred at the 90% confidence level. The 95% confidence intervals are not included but can be found in Table 4. Regions where lines are clustered together indicate a stronger agreement by multiple data sources that a significant change occurred in ILI percentage at this time.

For the 2012-2013 year, weeks 45, 50, and 51 of 2012 are the maximum peak week for Flu Near You, Athena, and ER Centers respectively (**Table 1**). Change point analysis results did not contain any change points from the Flu Near You data for this time interval. However, from the Athena data, change point analysis resulted in both an increase (2012 week 46; 95% CI: 2012 44th – 2012 49th) and decrease (2013 week 6; 95% CI; 2013 4th – 2013 8th) in average ILI. This indicates the duration of the influenza surge as approximately between 2012-11-11 and 2013-02-03. From the hospital ER centers, three increases and three decreases are identified. Change points from this source are identified earlier than from Athena and indicate a longer duration (**Table 4, Figure 4**).

For the 2013-2014 time period, Flu Near You, Athena, hospital ER centers, and ILINet show that the 51st week of 2013 is the peak week (**Table 1**). Using change point analysis, we were again not able to find any change points for FNY data. Athena data displayed a change point as early as the 29th week of 2013 (95% CI: 2013 29th -2013 29th). However, this is a decrease and therefore does not accurately estimate the start of an influenza surge. This change point also has the lowest confidence level (90%) of all recognized changes. From Athena, we do see a subsequent increasing change point at the 37th week of 2013 (95% CI: 2013 35th – 2013 38th) and the last decrease at the first week of 2014. The ER centers data also shows an early increase in the 35th week (95% CI: 2013 34th – 2013 42nd) but then a later and larger increase is identified in the 45th week of 2013 (95% CI: 2013 45th – 2013 46th). There is a large decrease in the 2nd week of 2014 (95% CI: 2014 1st – 2014 2nd) with a smaller, less meaningful decrease in the 20th week (95% CI: 2014 11th – 2014 20th). The first decrease detected is more reliable than the second due to the smaller confidence interval and higher confidence level. For this time

interval, ILINet data pinpoints increases in the 45th (95% CI: 2013 45th – 2013 45th) and 49th week of 2013 (95% CI: 2013 49th – 2013 51st). Both are meaningful changes that indicate the start of an influenza surge due to the narrow confidence intervals and high confidence levels. Both Athena and ER Centers data capture this increase in a similar period with the ER centers identifying the increase in the same week as ILINet. ILINet also detects a significant decrease in the 2nd week of 2014 (95% CI: 2014 2nd – 2014 2nd) with additional decreasing change points in the 11th (95% CI: 2014 11th -2014 11th) and 20th week (95% CI: 2014 18th- 2014 20th). Both Athena and ER Centers identify the initial decreasing change point close to the change point identified by ILINet. For the 2013-2014 season, both ER centers data and Athena could identify the beginning and ending of the influenza surge when compared to ILINet (**Table 4, Figure 4)**.

The maximum ILI activity peaks for 2014-2015 are the 37th (FNY), 50th (Athena), 51st (ER Centers), and 46th weeks of 2014 (**Table 1**). All data sources, including Flu Near You, identify significant change points in this season. Flu Near You detects a significant increase in the 47th week (95% CI: 2014 45th – 2014 48th) and a decrease in the 3rd week of 2015 (95% CI: 2015 2nd – 2015 5th). Athena also shows a decrease in the 3rd week of 2015 (95% CI: 2014 53rd – 2015 3rd), although the confidence interval is somewhat wide. The increase in Athena in the same week identified by Flu Near You, the 47th week of 2014 (95% CI: 2014 45th – 49th 2015), has a narrower interval. Athena also displays an additional, less significant increase (2015 10th; 95% CI: 2015 7th – 2015 20th) and decrease (2015 24th; 95% CI: 2015 15th – 2015 24th). The ER centers data shows three significant increases and six significant decreasing change points in this season. Like Flu Near You and Athena, the change point with the greatest increase is in the 46th week of

2014 (95% CI: 2014 46th – 2014 46th). A decrease with a narrow confidence interval is also identified in the 4th week of 2015 (95% CI: 2015 4th -2015 4th), like Athena and Flu Near You. Subsequent decreasing change points are also captured, describing the waning influenza activity. Change point analysis resulted in two increases and three decreases in the ILINet data for this season. Like the other sources, the 45th week of 2014 (95% CI: 2014 45th – 2014 45th) displays the narrowest confidence interval with the greatest increase in average ILI percentage. Similarly, ILINet also displayed a significant decrease in the 3rd week of 2015 (95% CI: 2015 3rd – 2015 3rd). For this time interval, all sources seem to capture similar increasing and decreasing change points (**Table 4, Figure 4**). There is variability in the number of change points identified by each data capture mechanism.

Weeks with the highest ILI percentage in the 2015 – 2016 season are the 25th (FNY), 12th (Athena), and 9th (ER Centers) weeks of 2016 (**Table 1**). There are no significant change points identified from the Flu Near You data again. The Athena data shows an increase in the 10th week of 2016 (95% CI: 2016 9th – 2016 12th) and a decrease in the 16th week (95% CI: 2016 14th – 2016 17th). Per the ER centers data, the greatest increase in average ILI percentage occurs in the 7th week of 2016 (95% CI: 2016 7th – 2016 8th). A subsequent decrease is identified in the 19th week (95% CI: 2016 18th – 2016 19th). We do see more gradual increases earlier in the 35th (95% CI: 2015 35th -2015 35th) and 43rd weeks (95% CI: 2015 40th – 2016 3rd) of 2015, though the latter change point shows a wide confidence interval. ILINet indicates an initial gradual increase as early as the 33rd week of 2015 (95% CI: 2015 32nd – 2015 39th) and the first decrease in the 6th week of 2016 (95% CI: 2016 5th – 2016 6th) which is comparatively earlier than the

change points identified from the other sources. The largest increase is identified in the 47th week of 2015 (95% CI: 2015 44th – 2015 47th). The ER centers data is the only one that shows a significant change as early as the 43rd week in a period comparable to ILINet, but as mentioned previously, the confidence interval for this change is very large. ILINet exhibits a decrease early in 2016 when Athena and ER Centers estimate that change to be an increase in ILI activity. Although the peak activity for influenza in this season may have been later and milder, ILINet is the only one that captures the beginning estimate of the typical influenza season **(Table 4, Figure 4)**.

Excluding Flu Near You, at least an increase and decrease are identified from all data capture mechanisms using change point analysis. The change point increases indicate that the 2015-2016 season had a later start for peak influenza activity. Overall, looking at changes in average ILI percentage, this season was also milder compared to previous seasons. The analysis also identifies other earlier and later significant change points that show the gradual increases and decreases of influenza activity.

	Table 4. Weeks (95% Confidence Interval) of Significant Change Points in ILI Percentages by Data Capture Mechanism (2012-2016)					
Weeks[1]	**Change Point**	**95% Confidence Interval**	**Average ILI % Prior to Change**	**Average ILI % After Change**	**Confidence Level[2]**	**Increase or Decrease**
2012-2013						
Flu Near You	No significant changes identified.					
Athena	2012 46th	2012 44th – 2012 49th	0.20	0.39	100%	⇧
	2013 6th	2013 4th – 2013 8th	0.39	0.19	100%	⇩
ER Centers	2012 35th	2012 35th – 2012 35th	0.76	1.34	100%	⇧
	2012 42nd	2012 42nd – 2012 42nd	1.34	2.41	91%	⇧

	2012 46th	2012 46th – 2012 46th	2.41	4.09	98%	⇧
	2013 4th	2013 4th – 2013 4th	4.09	2.14	100%	⇩
	2013 11th	2013 11th – 2013 11th	2.14	1.46	100%	⇩
	2013 19th	2013 19th – 2013 19th	1.46	1.04	100%	⇩
ILINet	36th 2012	36th 2012 – 43rd 2012	0.16	0.66	94%	⇧
	48th 2012	48th 2012 – 48th 2012	0.66	2.20	100%	⇧
	5th 2013	5th 2013 – 5th 2013	2.20	1.05	99%	⇩
	10th 2013	9th 2013 – 10th 2013	1.05	0.41	95%	⇩
	15th 2013	13th 2015 – 15th 2013	0.41	0.15	98%	⇩
	22nd 2013	22nd 2013 – 22nd 2013	0.15	0.06	98%	⇩
2013-2014						
Flu Near You	No significant changes identified.					
Athena	2013 29th	2013 29th – 2013 29th	0.24	0.09	90%	⇩
	2013 37th	2013 35th – 2013 38th	0.09	0.17	100%	⇧
	2013 47th	2013 47th – 2013 51st	0.17	0.34	95%	⇧
	2014 1st	2013 51st – 2014 1st	0.34	0.11	100%	⇩
ER Centers	2013 35th	2013 34th– 2013 42nd	0.74	1.49	100%	⇧
	2013 45th	2013 45th – 2013 46th	1.49	4.53	100%	⇧
	2014 2nd	2014 1st – 2014 2nd	4.53	1.58	100%	⇩
	2014 20th	2014 11th – 2014 20th	1.58	1.06	97%	⇩
ILINet	2013 45th	2013 45th – 2013 45th	0.17	1.72	100%	⇧
	2013 49th	2013 49th – 2013 51st	1.72	3.05	95%	⇧
	2014 2nd	2014 2nd – 2014 2nd	3.05	0.78	100%	⇩

	2014 11th	2014 11th – 2014 11th	0.78	0.23	100%	⇩
	2014 20th	2014 18th – 2014 20th	0.23	0.05	100%	⇩
2014-2015						
Flu Near You	2014 47th	2014 45th – 2014 48th	1.52	7.67	100%	⇧
	2015 3rd	2015 2nd – 2015 5th	7.67	1.98	100%	⇩
Athena	2014 47th	2014 45th – 2014 49th	0.06	0.16	100%	⇧
	2015 3rd	2014 53rd – 2015 3rd	0.16	0.07	98%	⇩
	2015 10th	2015 7th – 2015 20th	0.07	0.11	92%	⇧
	2015 24th	2015 15th – 2015 24th	0.11	0.05	93%	⇩
ER Centers	2014 28th	2014 28th – 2014 30th	0.63	0.53	91%	⇩
	2014 35th	2014 35th – 2014 36th	0.53	1.02	96%	⇧
	2014 41st	2014 41st – 2014 41st	1.02	1.84	98%	⇧
	2014 46th	2014 46th – 2014 46th	1.84	3.47	96%	⇧
	2014 53rd	2014 53rd – 2014 53rd	3.47	2.60	100%	⇩
	2015 4th	2015 4th – 2015 4th	2.60	1.89	97%	⇩
	2015 11th	2015 11th – 2015 11th	1.89	1.43	97%	⇩
	2015 15th	2015 15th – 2015 15th	1.43	0.88	95%	⇩
	2015 21st	2015 21st – 2015 21st	0.88	0.57	99%	⇩
ILINet	2014 36th	2014 36th - 2014 44th	0.02	0.19	99%	⇧
	2014 45th	2014 45th – 2014 45th	0.19	1.34	100%	⇧
	2014 50th	2014 50th – 2014 50th	1.34	0.71	99%	⇩
	2015 3rd	2015 3rd – 2015 3rd	0.71	0.20	100%	⇩
	2015 12th	2015 8th – 2015 12th	0.20	0.03	99%	⇩

2015-2016						
Flu Near You	No significant changes identified.					
Athena	2016 10th	2016 9th – 2016 12th	0.10	0.28	100%	⇧
	2016 16th	2016 14th – 2016 17th	0.28	0.13	97%	⇩
ER Centers	2015 35th	2015 35th – 2015 35th	0.40	0.97	100%	⇧
	2015 43rd	2015 40th – 2016 3rd	0.97	1.18	91%	⇧
	2016 7th	2016 7th – 2016 8th	1.18	2.14	100%	⇧
	2016 19th	2016 18th – 2016 19th	2.14	0.99	100%	⇩
ILINet[3]	33rd 2015	32nd 2015 – 39th 2015	0.03	0.14	93%	⇧
	47th 2015	44th 2015 – 47th 2015	0.14	0.53	100%	⇧
	6th 2016	5th 2016 – 6th 2016	0.53	0.18	100%	⇩
	12th 2016	11th 2016 – 12th 2016	0.18	0.06	99%	⇩
	15th 2016	15th 2016 – 15th 2016	0.06	0.01	99%	⇩
	26th 2016	21st 2016 – 26th 2016	0.01	0.03	93%	⇧

[1]Time period is week 27 from initial year through week 26 of follow year.
[2]Must meet at least 90% confidence level for inclusion in table.
[3] ILINet data in this time interval contained one or more outliers. Custom analysis on the ranks of values was performed using Change-Point Analyzer.

Forecasting Future Influenza Activity Trends from ARIMA

From the available ILI percentage data, we find the best fitting ARIMA models (**Table 5**). Using the forecast ARIMA equations, we estimate ILI percentage point estimates and 95% confidence interval for the next ten weeks, weeks 27 through 36 of 2016, for each data source. The Athena model results in the same ten point estimates of 0.12% and 95% confidence intervals are wide and cross zero, but magnitude increases

steadily with mean confidence range of 0.35%. Flu Near You point estimates continually decrease from 3.26% (95% CI: -1.00 – 7.51) to 2.06% (95% CI: -2.59 – 6.72). The 95% confidence intervals for Flu Near You point estimates also cross zero and are wide with a mean confidence range of 9.18%. The ILI percentage estimates for ER Centers continually increase from 0.72% (95% CI: 0.19 – 1.26) to 1.32% (95% CI: -0.35 – 3.00). All except the initial confidence interval cross zero and the mean range is 2.43%. Lastly, ILINet ILI percentage estimates start at 0.01% (95% CI: -0.23 – 0.25), decrease to -0.09% (95% CI: -0.70 – 0.53), and increase slowly again to -0.05% (95% CI: 1.03 – 0.92). The mean range between upper and lower 95% limits was 1.18%. Like the other data sources, the 95% confidence interval also crosses zero for ILINet.

The Athena data source has the lowest average confidence interval difference between the upper and lower 95% bounds (0.35%). Flu Near You has the highest (9.18%). All the 95% confidence intervals are relatively wide and lower bounds are negative percentages, which cannot be actualized because ILI percentages can only be positive. The data sources do not agree on an increasing or decreasing predictive trend (**Figure 5**).

Discussion

Strengths

Because there are four years of historical data available, the analysis is less affected by a few weeks with a small number of observations. The first five weeks of Flu Near You contain less than five total responses. When comparing the distribution with and without these values, a meaningful or significant difference is not found. The sample size is large enough and that smaller changes in weekly ILI activity do not tilt the data to show unreliable predictions.

The varying data sources also exhibit similar patterns when it comes to seasonality and peaks based on descriptive analysis. The results from change point analysis also show similar significant increasing and decreasing change points (**Table 4**). The overlap of the different sources demonstrates a greater cohesion and more informative results when considered simultaneously. Each may be limited on its own, but comparing all different data capture mechanisms can give an earlier, richer picture of influenza activity.

The turn-over time is also faster when considering various data sources instead of only influenza laboratory results. When using user-reported data in conjunction with automated systems, we can see minute changes faster than traditional systems that may require multiple processes for influenza confirmation. The data sources evaluated here are best utilized together.

The data capture mechanisms cover varying stages of influenza. Google Flu Trends captures users who searched influenza-related terms. This source had the potential to capture those who may not have seen a provider yet but are curious about their

symptoms or potential treatments. Flu Near You reporters may not have a confirmed diagnosis but input their responses based on changes in their symptoms weekly. Reports from the Athena Network capture visits from patients who perceive themselves to be sick enough to visit their provider but are recorded as outpatient because of lower severity. The hospital ER centers collect information from patients who perceive their symptoms to be so severe that they need immediate attention. Using various data types allows a more comprehensive view of influenza trends with the potential for earlier detection.

Limitations

ILINet is considered the gold standard for this analysis. Ideally, we would have liked to validate results by comparing data to influenza laboratory testing results. The influenza laboratory test data is too sparse to use as a gold standard to accurately compare to the other data sources. Because availability of specimens for influenza testing was low, this data source was not included in the analysis. Future analysis would benefit from comparison to influenza laboratory testing to gain a more accurate result of actual influenza activity in the population.

Flu Near You users report on their symptoms every Monday. One week for this data source is considered Monday-Sunday. However, all other data sources consider a week to be Sunday - Saturday. Since six out of seven days do overlap, to be able to compare all sources with their corresponding weeks, we will assume that the one day difference in the Flu Near You weeks does not alter the results. We assume that this will not cause significant changes to the accuracy of comparisons. Published analyses comparing ILINet and Flu Near You also showed that this would not significantly alter results (Bakota, Santos, & Arafat, 2015; Smolinski et al., 2015). There were very few

users during the initial stages and inconsistency and variability in the number of users who self-reported using Flu Near You during the initial stages is high. The Flu Near You data does decrease in variability as the number of users reporting increase.

GFT was a controversial program that tried to predict influenza activity using search queries. It was largely criticized for inaccurate magnitudes, especially during the 2009 H1N1 event. Only weekly ILI count available for this analysis. Since this source deals with search queries made through Google, it would be difficult to obtain the total number of search queries in a week as this number would be very large. Although we can get population information for the City of Houston, we cannot estimate the amount of search queries a person makes or that this rate is uniform. This source is included when comparing ILI counts but excluded from further analysis where ILI percentage will be more informative. There is no data for the 2015-2016 influenza season for GFT.

Change point analysis requires an independent error assumption. However, some of the data capture mechanisms violated this assumption for some of the time intervals. Because of the nature of this type of data, one can expect positive correlation of errors or that during influenza surges, if one value is above average, the next several values will also tend to be above average until the duration of the epidemic or outbreak diminishes. The opposite can also be true resulting in negative correlation of errors. Autocorrelation does not significantly impact change point analysis performance and this has also been evidenced in a previous article using ILI emergency department visits. According to this article, the CUSUM CPA method is still robust enough to detect change points with autocorrelation in time series data (Kass-Hout et al., 2012).

Forecasting time series data is difficult due to inherent uncertainties of trend and retention of historical properties. Our predictions are based on historical data and would not be ideal to predict unusual events that have not occurred previously. Forecasting also must consider the noise component which is difficult to estimate or predict. It is unclear to what degree noise affects the future model. However, taking into the degree of uncertainty, short-term estimates can be useful for preparation or increased level of awareness.

Further Considerations

Analyses is conducted on each data source separately to compare results to each other to see if there is agreement of influenza peaks, trends, and magnitude. Piecing together combines the results together the complementary corresponding information. To validate the forecasted values using ARIMA, we will compare results to actual values observed during these ten weeks. This was not completed during this analysis because this data was not available at this time. Further analyses are needed to assess if these sources can be combined to reliably identify significant points of interest in the influenza season early and to accurately forecast estimates of future influenza activity. A single result from such analysis would be more objective when using all the available information.

Conclusion

Overall, the outpatient data from the Athena Network, emergency center data from the Houston ER Centers, and ILINet data show similar trends. The identified change points exhibit similar increase and decrease in ILI activity when it comes to identifying the start and end of the influenza season (**Table 4**). The peaks of ILI percentages from these sources and Flu Near You are relatively close to each other (**Table 1**). The ARIMA forecast estimates produce different trends regarding predicting the next ten weeks of event activity. Combining the information from the diverse data sources can identify with more certainty the start and end of influenza activity and return improved forecast estimates.

For agencies that are not able to access multiple sources of ILI surveillance information, when considering consistency in the number of observations, ER Centers data and ILINet are likely most useful in influenza surveillance. Athena can also be a useful source, but Flu Near You usefulness in a specific community will depend on the consistency with which users in that community report their symptoms. Flu Near You will also need a large enough sample size to offer reliable estimates from its data. If it meets these two requirements, this is a tool that can quickly estimate influenza or influenza-like activity. If only investing in one of these influenza-like illness surveillance sources, ILINet or ER Centers data would likely give the most information.

The identification of similar significant points of increase and decrease in the influenza season assist public health professionals in preparing for potential surges of influenza activity. Because there are similarities between the different weeks identified as points of interest, we can conclude that the sources are precise. If surveillance

professionals look at all available types of data that can capture information related to influenza, the health care community can ensure it is ready for any potential increases in vaccination demand, information demand, medication, hospitalizations, or other outcomes. Once more consistent and sufficient information is collected and analyzed, forecast estimates of influenza can enhance preparation and prevention efforts.

Tables and Figures

Table 3. Peak Weeks per Year and ILI Counts by Data Capture Mechanism

	Data Source	Peak Week [a]	Peak ILI Count	Mean	Standard Deviation
2012-2013 [c]	Flu Near You [b]	1st 2013	29.00	2.00	5.07
	Google Flu Trends	3rd 2013	12,069.00	4,118.71	2,782.40
	HHD ER Center	51st 2012	2329.00	7,201.35	484.78
	Athena Network	50th 2012	25.00	8.10	4.59
	ILINet	1st 2013	358.00	107.85	103.85
2013-2014 [c]	Flu Near You [b]	51st 2013, 2nd 2014	7.00	1.38	1.65
	Google Flu Trends	52nd 2013	9,617.00	2710.23	1,788.92
	HHD ER Center	52nd 2013	2,455.00	603.38	589.39
	Athena Network	51st 2013	26.00	6.83	4.63
	ILINet	50th 2013	469.00	78.94	107.05
2014-2015 [c]	Flu Near You [b]	47th 2014, 52nd 2014	11.00	2.30	2.97
	Google Flu Trends	52nd 2014	7,676.00	7,676.00	1,630.13
	HHD ER Center	51st 2014, 52nd 2014	1,546.00	586.02	397.45
	Athena Network	50th 2014, 16th 2015	15.00	6.13	3.58
	ILINet	47th 2014	200.00	37.09	54.07
2015-2016 [c]	Flu Near You [b]	9th 2016	10.00	1.92	1.85
	Google Flu Trends	Data unavailable.			
	HHD ER Center	10th 2016	463.00	203.00	98.56
	Athena Network	15th 2016	22.00	8.81	4.74
	ILINet	51st 2015	242.00	45.26	57.44

[a] Peak week is the week with the highest ILI count.
[b] Flu Near You data ranges from August 2012-current June 2016
[c] All year intervals begin MMWR week 27 of 2012 through MMWR week 26 of the following year.
[d] Google Flu Trends data included ranges from August 2011-July 2015.

Table 2. Stationarity and Seasonal Differencing Test Results				
Data Source	**KPSS p-value[3]**	**ADF p-value[2]**	**Stationary Differencing Required**	**Seasonal Differencing Required**
Athena[1]	<0.01	0.08	Yes	No
Flu Near You	0.1	<0.01	No	No
ER Centers	0.09	<0.01	No	No
ILINet	<0.01	<0.01	Yes	No

[1] First-order differencing will be required.
[2] P-values less than 0.05 indicate stationarity.
[3] P-values greater than 0.05 indicate stationarity.

Table 5. ARIMA Model Orders and Parameters		
Data Source	**Orders**	**Ljung-Box Test p-value[2]**
Athena	(0,1,1)	0.23
Flu Near You	(2,0,0)	0.59
ER Centers	(2,0,1)(1,0,0)[52]	<0.01[1]
ILINet	(2,1,1)(0,0,1)[52]	0.15

[1] These residuals are not random but because there is sufficient number of observations, this is the best model possible.
[2] P-values above 0.05 are desirable and indicate higher possibility of zero autocorrelation of residuals.

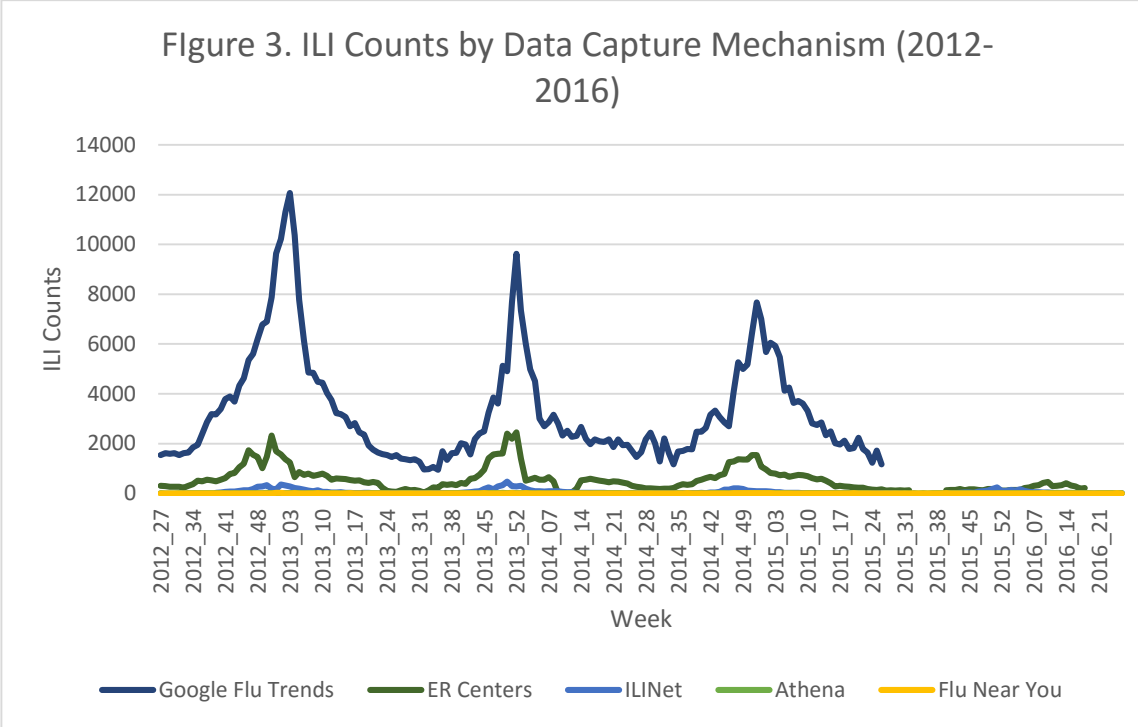
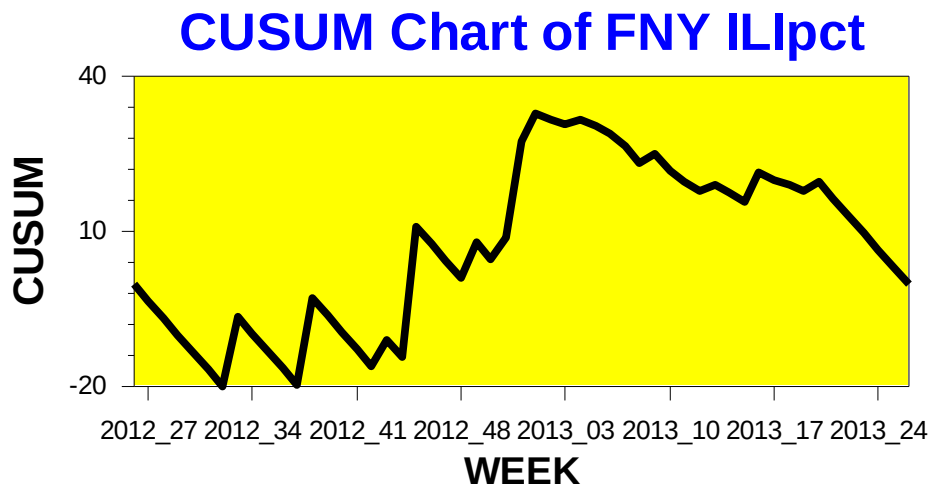
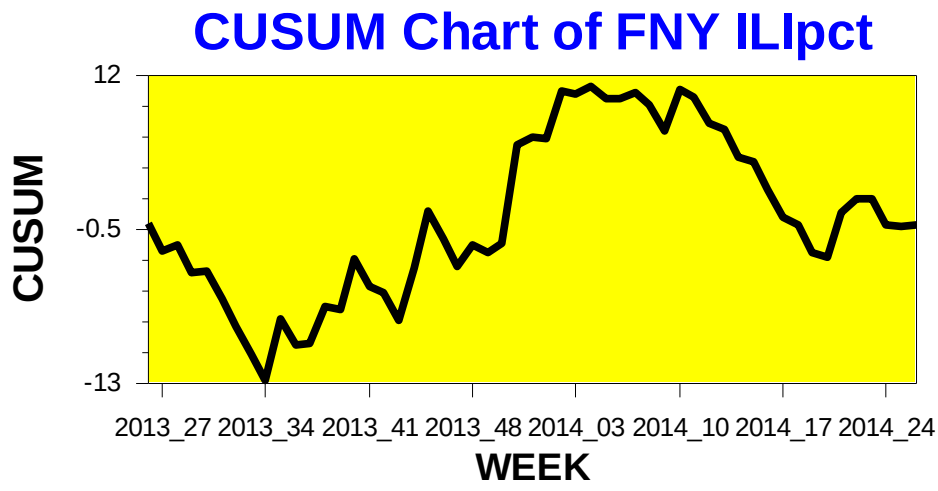
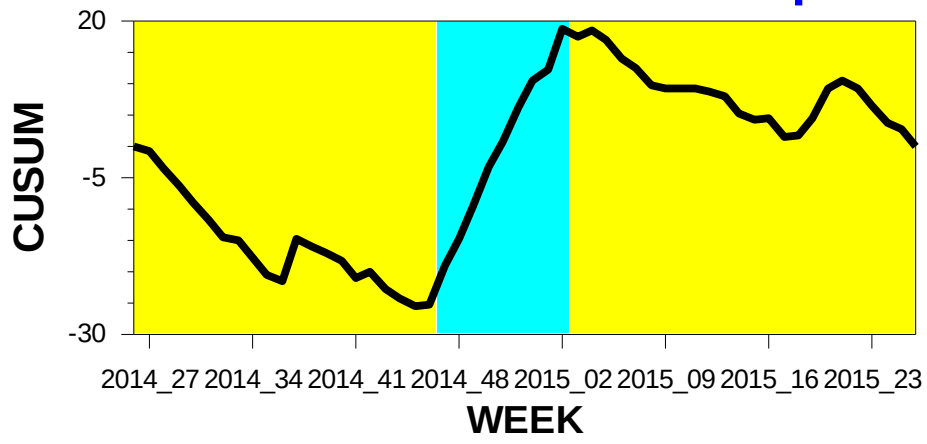


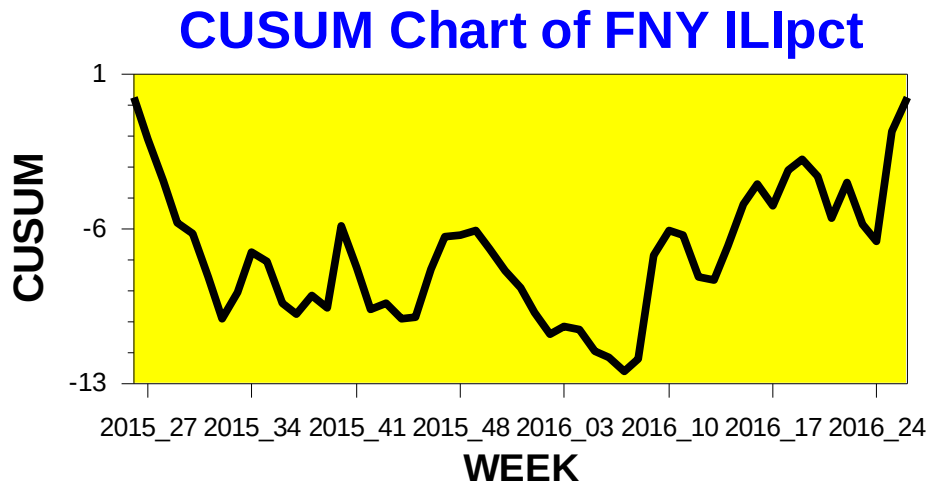
Figure 3. The above figure shows ILI counts by data capture mechanism from MMWR week 27 of 2012 through MMWR week 26 of 2016. Google Flu Trends shows the highest number of counts (search queries). Athena outpatient data and Flu Near You have the lowest ILI counts. Google Flu Trends, hospital ER centers, and ILINet show clear seasonal peaks at similar times. Any outliers or missing data was removed from the data.

Figure 2. CUSUM Charts of ILI Percentages (Week 27 2012 – Week 26 2016)
Flu Near You

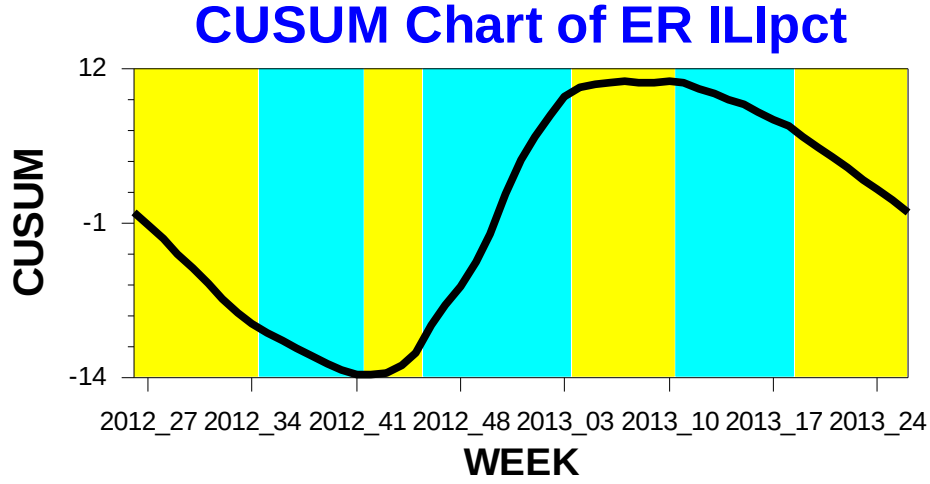


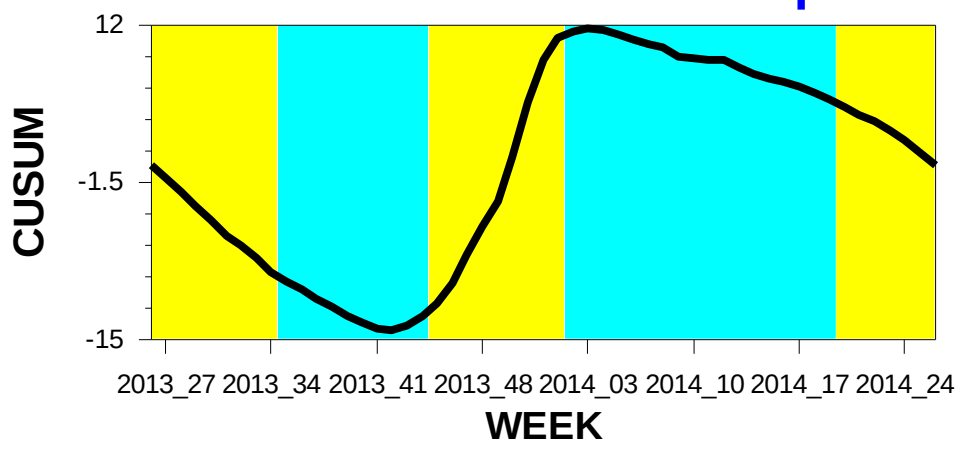


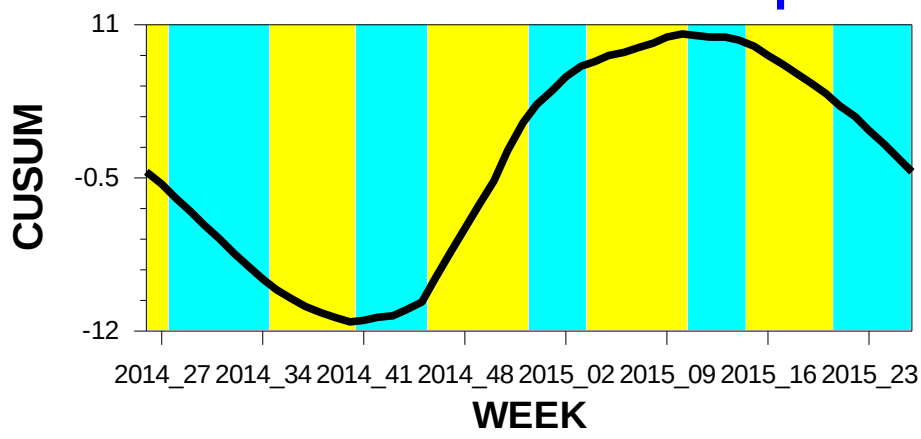

CUSUM Chart of FNY ILIpct
CUSUM
20
-5
-30
2014_27 2014_34 2014_41 2014_48 2015_02 2015_09 2015_16 2015_23
WEEK

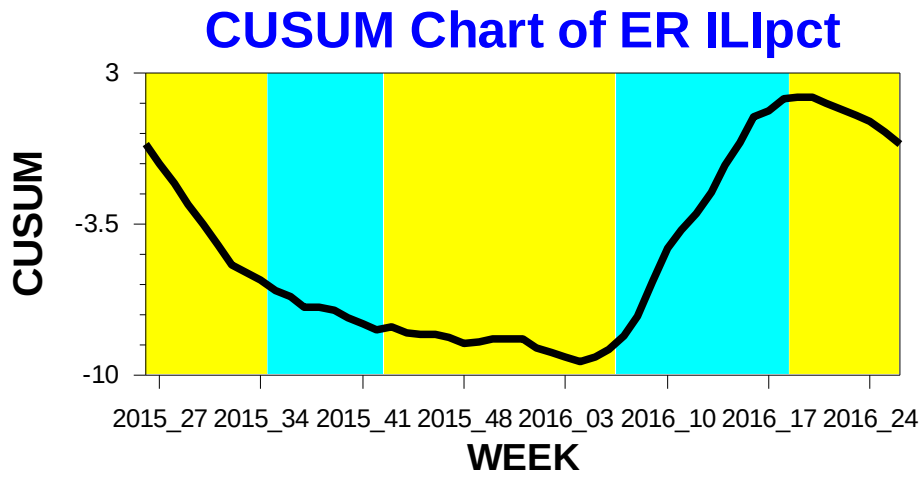

CUSUM Chart of FNY ILIpct
CUSUM
1
-6
-13
2015_27 2015_34 2015_41 2015_48 2016_03 2016_10 2016_17 2016_24
WEEK

ER Centers

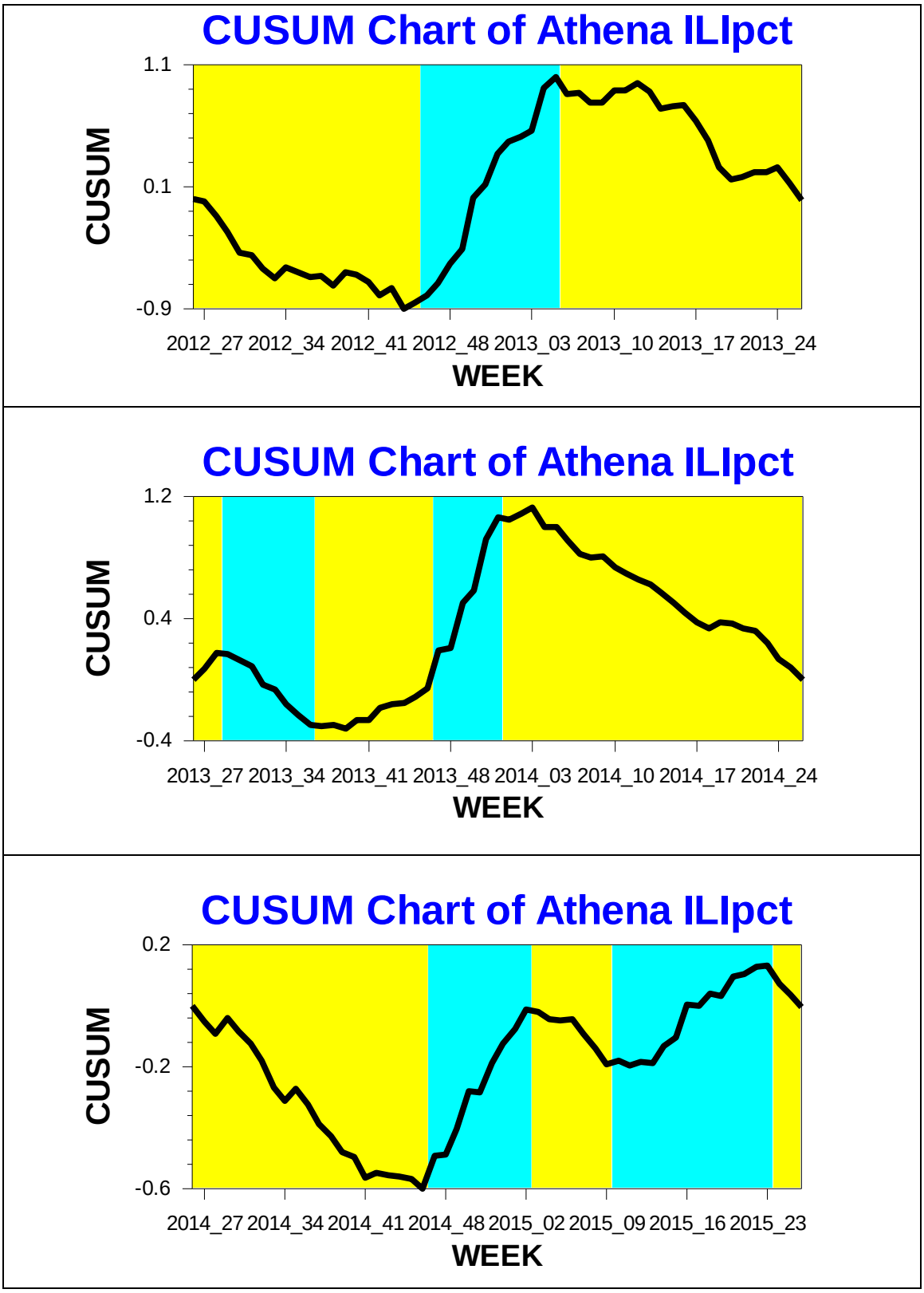

CUSUM Chart of ER ILIpct
CUSUM
12
-1
-14
2012_27 2012_34 2012_41 2012_48 2013_03 2013_10 2013_17 2013_24
WEEK

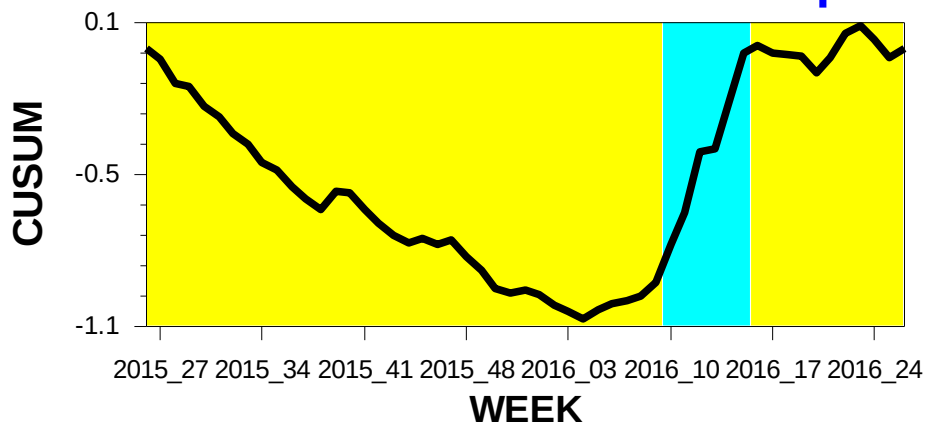

CUSUM Chart of ER ILIpct
CUSUM
12
-1.5
-15
2013_27 2013_34 2013_41 2013_48 2014_03 2014_10 2014_17 2014_24
WEEK


CUSUM Chart of ER ILIpct
CUSUM
11
-0.5
-12
2014_27 2014_34 2014_41 2014_48 2015_02 2015_09 2015_16 2015_23
WEEK

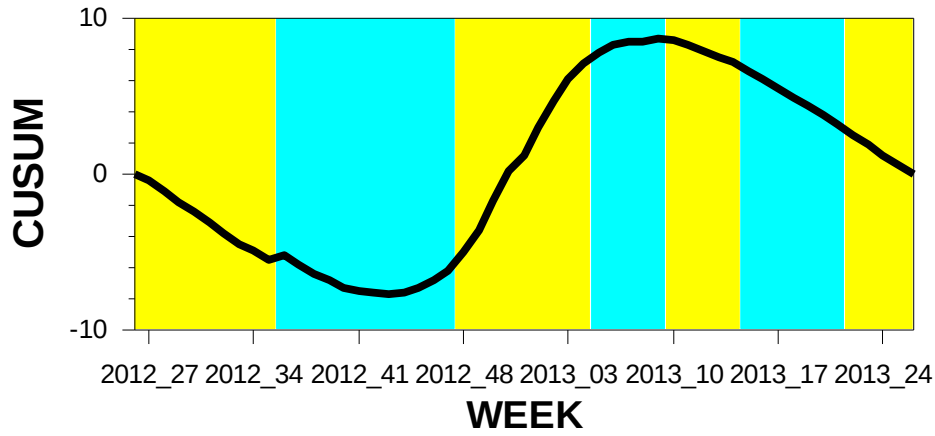

CUSUM Chart of ER ILIpct
CUSUM
3
-3.5
-10
2015_27 2015_34 2015_41 2015_48 2016_03 2016_10 2016_17 2016_24
WEEK

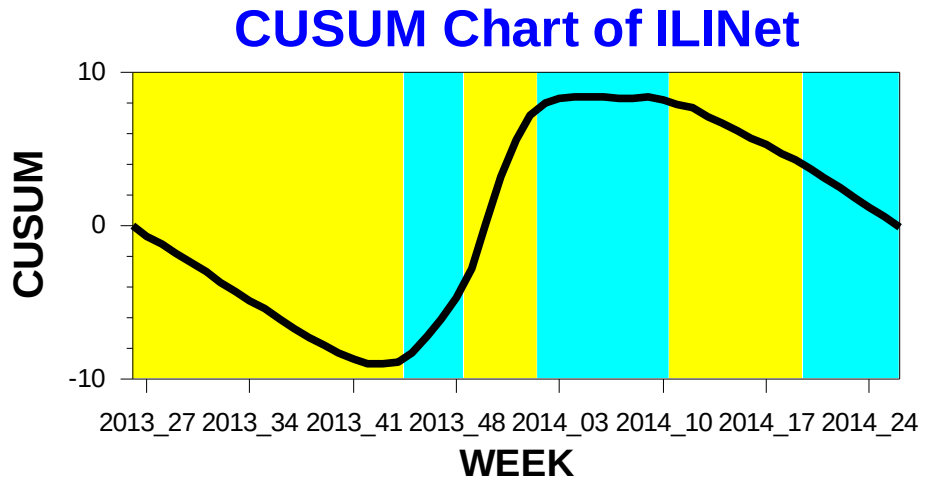
Athena


CUSUM Chart of Athena ILIpct
CUSUM
1.1
0.1
-0.9
2012_27 2012_34 2012_41 2012_48 2013_03 2013_10 2013_17 2013_24
WEEK
CUSUM Chart of Athena ILIpct
CUSUM
1.2
0.4
-0.4
2013_27 2013_34 2013_41 2013_48 2014_03 2014_10 2014_17 2014_24
WEEK
CUSUM Chart of Athena ILIpct
CUSUM
0.2
-0.2
-0.6
2014_27 2014_34 2014_41 2014_48 2015_02 2015_09 2015_16 2015_23
WEEK


CUSUM Chart of Athena ILIpct
CUSUM
0.1
-0.5
-1.1
2015_27 2015_34 2015_41 2015_48 2016_03 2016_10 2016_17 2016_24
WEEK

ILINet


CUSUM Chart of ILINet
CUSUM
10
0
-10
2012_27 2012_34 2012_41 2012_48 2013_03 2013_10 2013_17 2013_24
WEEK


CUSUM Chart of ILINet
CUSUM
10
0
-10
2013_27 2013_34 2013_41 2013_48 2014_03 2014_10 2014_17 2014_24
WEEK

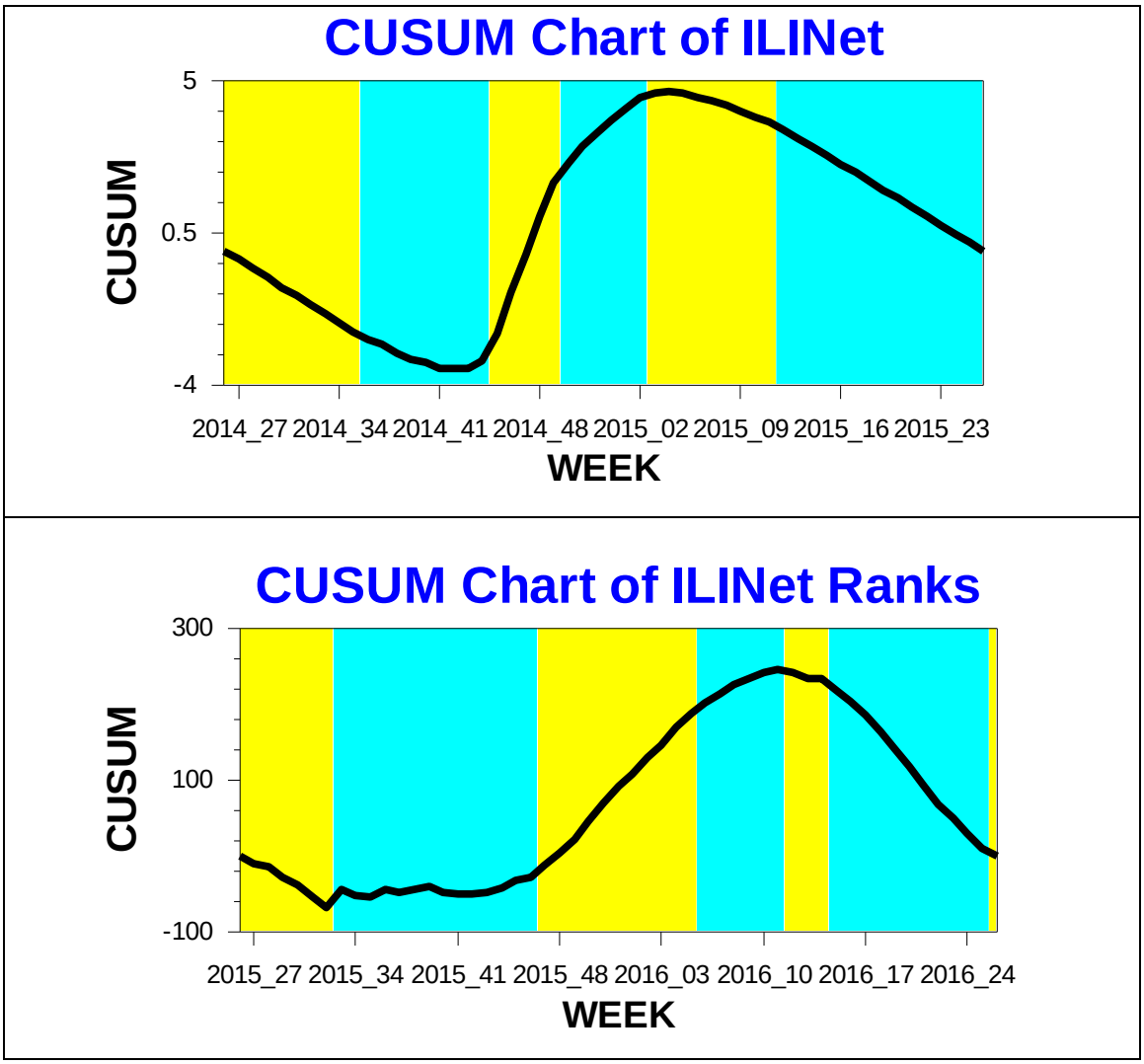
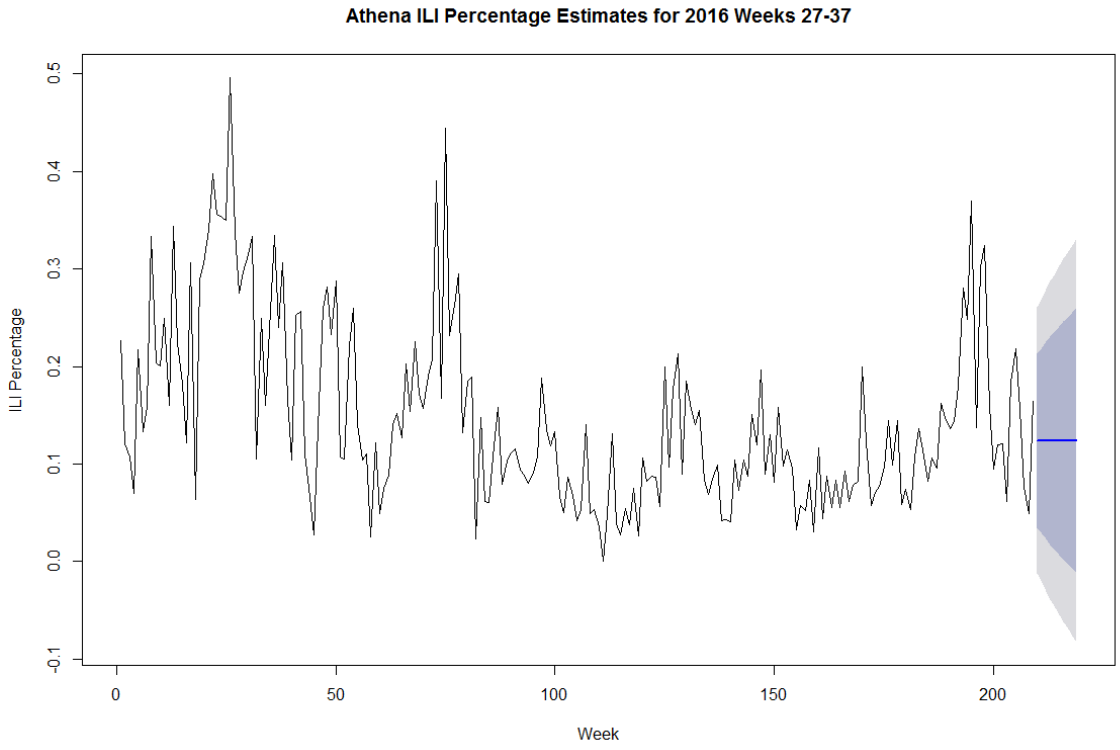
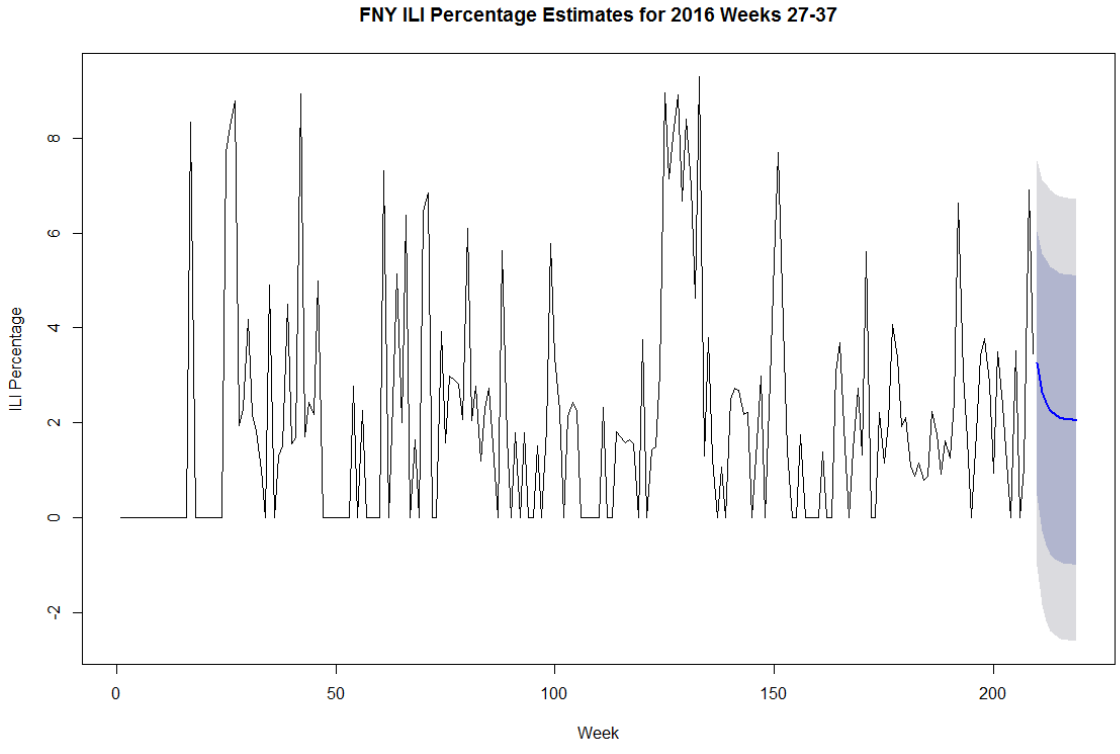

CUSUM Chart of ILINet
CUSUM
5
0.5
-4
2014_27 2014_34 2014_41 2014_48 2015_02 2015_09 2015_16 2015_23
WEEK
CUSUM Chart of ILINet Ranks
CUSUM
300
100
-100
2015_27 2015_34 2015_41 2015_48 2016_03 2016_10 2016_17 2016_24
WEEK

Figure 5. Point Estimates and 80% and 95% Confidence Intervals of ILI Percentages





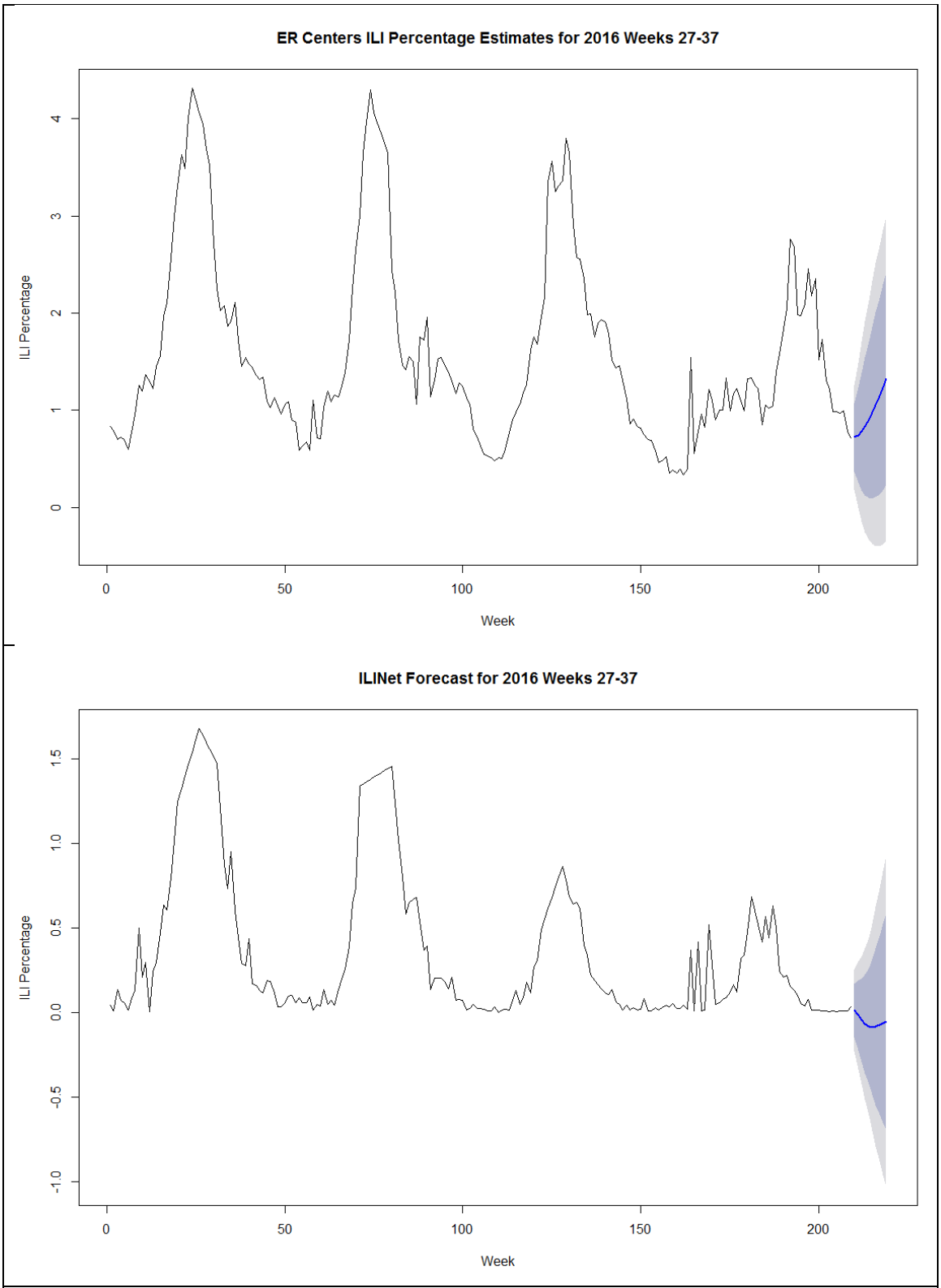


Figure 5. Each of these plots show the original data with the point estimates as a blue line and the confidence intervals extending out ten weeks. The darker bands represent

the 80% confidence intervals and the lighter bands represent the 95% confidence intervals. All intervals cross over the zero line and are wide.

References

Allison, P. (2015). Imputation by Predictive Mean Matching: Promise & Peril. *Statistical Horizons*. Retrieved from http://statisticalhorizons.com/predictive-mean-matching

Baddour, Y., Tholmer, R., & Gavit, P. (2009). Use of Change-Point Analysis for Process Monitoring and Control. *BioPharm International*, *22*(8). Retrieved from http://www.biopharminternational.com/use-change-point-analysis-process-monitoring-and-control

Bakota, E. V, Santos, E. R., & Arafat, R. R. (2015). Estimating FluNearYou Correlation to ILINet at Different Levels of Participation, *8*(1), 2579.

Caspard, H., Gaglani, M., Clipper, L., Belongia, E. A., McLean, H. Q., Griffin, M. R., … Ambrose, C. S. (2016). Effectiveness of live attenuated influenza vaccine and inactivated influenza vaccine in children 2-17 years of age in 2013-2014 in the United States. *Vaccine*, *34*(1), 77–82. https://doi.org/10.1016/j.vaccine.2015.11.010

Chaves, S. S., Perez, A., Farley, M. M., Miller, L., Schaffner, W., Lindegren, M. L., … Finelli, L. (2014). The Burden of Influenza Hospitalizations in Infants from 2003-2012, United States. *The Pediatric Infectious Disease Journal*, *33*(9), 912–919. https://doi.org/10.1097/INF.0000000000000321

Collection, S. (2016). HHS Public Access, *8*(5), 583–592. https://doi.org/10.1002/aur.1474.Replication

Davidson, M. W., Haim, D. A., & Radin, J. M. (2015). Using Networks to Combine "Big Data" and Traditional Surveillance to Improve Influenza Predictions. *Scientific Reports*, *5*(Region 3), 8154. https://doi.org/10.1038/srep08154

Dawood, F. S., Iuliano, A. D., Reed, C., Meltzer, M. I., Shay, D. K., Cheng, P. Y., …

Widdowson, M. A. (2012). Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: A modelling study. *The Lancet Infectious Diseases*, *12*(9), 687–695. https://doi.org/10.1016/S1473-3099(12)70121-4

Dugas, A. F., Jalalpour, M., Gel, Y., Levin, S., Torcaso, F., Igusa, T., & Rothman, R. E. (2013). Influenza Forecasting with Google Flu Trends. *PLoS ONE*, *8*(2). https://doi.org/10.1371/journal.pone.0056176

Fowlkes, A., Dasgupta, S., Chao, E., Lemmings, J., Goodin, K., Harris, M., … Finelli, L. (2013). Estimating influenza incidence and rates of influenza-like illness in the outpatient setting. *Influenza and Other Respiratory Viruses*, *7*(5), 694–700. https://doi.org/10.1111/irv.12014

Hay, S. I., George, D. B., Moyes, C. L., & Brownstein, J. S. (2013). Big Data Opportunities for Global Infectious Disease Surveillance. *PLoS Medicine*, *10*(4). https://doi.org/10.1371/journal.pmed.1001413

Henning, K. J. (New Y. C. D. of H. and M. H. (2004). Overview of Syndromic Surveillance What is Syndromic Surveillance? *MMWR*, *53*(Suppl), 5–11. Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/su5301a3.htm#top

Hiller, K. M., Stoneking, L., Min, A., & Rhodes, S. M. (2013). Syndromic Surveillance for Influenza in the Emergency Department-A Systematic Review. *PLoS ONE*, *8*(9), 6–10. https://doi.org/10.1371/journal.pone.0073832

Hyndman, R. J., & Athanasopoulos, G. (2014). *Forecasting: principles and practice*. Otexts.com. Retrieved from https://www.otexts.org/fpp/8

Kass-Hout, T. a, Xu, Z., McMurray, P., Park, S., Buckeridge, D. L., Brownstein, J. S., …

Groseclose, S. L. (2012). Application of change point analysis to daily influenza-like illness emergency department visits. *Journal of the American Medical Informatics Association : JAMIA*, *19*(6), 1075–1081. https://doi.org/10.1136/amiajnl-2011-000793

Khan, A. S., & Levitt, A. M. (2000). *Biological and Chemical Terrorism: Strategic Plan for Preparedness and Response*. Retrieved from https://www.cdc.gov/mmwr/preview/mmwrhtml/rr4904a1.htm

Rosenkötter, N., Ziemann, A., Riesgo, L. G.-C., Gillet, J. B., Vergeiner, G., Krafft, T., & Brand, H. (2013). Validity and timeliness of syndromic influenza surveillance during the autumn/winter wave of A (H1N1) influenza 2009: results of emergency medical dispatch, ambulance and emergency department data from three European regions. *BMC Public Health*, *13*, 905. https://doi.org/10.1186/1471-2458-13-905

Sagiroglu, S., & Sinanc, D. (2013). Big data: A review. In *2013 International Conference on Collaboration Technologies and Systems (CTS)* (pp. 42–47). https://doi.org/10.1109/CTS.2013.6567202

Seo, D. W., Jo, M. W., Sohn, C. H., Shin, S. Y., Lee, J. H., Yu, M., … Lee, S. Il. (2014). Cumulative query method for influenza surveillance using search engine data. *Journal of Medical Internet Research*, *16*(12), 1–16. https://doi.org/10.2196/jmir.3680

Shaman, J., & Karspeck, A. (2012). Forecasting seasonal outbreaks of influenza. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(3), 20425–30. https://doi.org/10.1073/pnas.1208772109

Skowronski, D. M., Janjua, N. Z., De Serres, G., Sabaiduc, S., Eshaghi, A., Dickinson, J.

A., … Li, Y. (2014). Low 2012-13 influenza vaccine effectiveness associated with mutation in the egg-adapted H3N2 vaccine strain not antigenic drift in circulating viruses. *PLoS ONE*. https://doi.org/10.1371/journal.pone.0092153

Smith, K. J., Raviotta, J. M., Depasse, J. V., Brown, S. T., Shim, E., Patricia Nowalk, M., & Zimmerman, R. K. (2016). Cost effectiveness of influenza vaccine choices in children aged 2-8 years in the U.S. *American Journal of Preventive Medicine*, *50*(5), 600–608. https://doi.org/10.1016/j.amepre.2015.12.010

Smolinski, M. S., Crawley, A. W., Baltrusaitis, K., Chunara, R., Olsen, J. M., Wójcik, O., … Brownstein, J. S. (2015). Flu near you: Crowdsourced symptom reporting spanning 2 influenza seasons. *American Journal of Public Health*, *105*(10), 2124–2130. https://doi.org/10.2105/AJPH.2015.302696

Suh, M., Kang, D. R., Lee, D. H., Choi, Y. J., Tchoe, B., Nam, C. M., … Kim, C. (2013). Socioeconomic burden of influenza in the Republic of Korea, 2007-2010. *PLoS ONE*, *8*(12), 2007–2010. https://doi.org/10.1371/journal.pone.0084121

Taylor, W. A. (2000). Change-Point Analysis: A Powerful New Tool For Detecting Changes. Retrieved from http://www.variation.com/cpa/tech/changepoint.html

Viboud, C., Miller, M. A., Grenfell, B. T., Bj??rnstad, O. N., & Simonsen, L. (2007). Erratum: Air travel and the spread of influenza: Important caveats (PLoS Medicine 3:11 doi: 10.1371/journal.pmed.0030503). *PLoS Medicine*. https://doi.org/10.1371/journal.pmed.0040032

Yang, W., Cowling, B. J., Lau, E. H. Y., & Shaman, J. (2015). Forecasting Influenza Epidemics in Hong Kong. *PLoS Computational Biology*, *11*(7). https://doi.org/10.1371/journal.pcbi.1004383