SURFING FOR FLU: COMPARING GOOGLE, TWITTER, AND WIKIPEDIA AS TOOLS FOR INFLUENZA SURVEILLANCE USING BAYESIAN CHANGE POINT ANALYSIS

By

JOYCE DANIELLE SHARPE

A THESIS PRESENTED TO THE GRADUATE SCHOOL OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

UNIVERSITY OF FLORIDA

2016

To my loving family, dearest friends, and trusted mentors who have contributed to my intellectual, professional, and personal growth

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my thesis chair, Dr. Robert L. Cook, and my committee members, Dr. Catherine W. Striley and Dr. Richard S. Hopkins, for their valuable insight and continuous guidance. I wish to thank Dr. Cook for his encouragement and support throughout this process and my graduate studies at the University of Florida. I would like to thank Dr. Striley for her consistent mentorship, sharp insight, and facilitation of my professional development. I would like to extend special appreciation to Dr. Hopkins for providing direction, wisdom, and expertise throughout this process and challenging me to truly evaluate the practicality of my research.

I would like to thank Dr. Heather Rubino from the Florida Department of Health as well as Sophie Smith and Matthew Biggerstaff from the Centers for Disease Control and Prevention for their insight, expertise, and time in regards to discussing influenza surveillance. I extend gratitude to Dr. John Emerson from Yale University, Dr. Chandra Erdman from Management Leadership for Tomorrow, and Dr. Xiaofei Wang from Amherst College for their assistance with understanding the complex statistics behind the methods used in this study. I am appreciative of Google, Inc., Dr. Mark Dredze and his team at HealthTweets.org, and Wikimedia Statistics for making their data publicly accessible.

I would also like to express thanks to my closest friends for their unwavering encouragement, emotional and spiritual support, and friendship.

My heartfelt gratitude goes to my wonderful sisters, Sophia Sharpe, Jennifer Sharpe, and Olivia Sharpe, for keeping me laughing and grounded when I needed to be

reminded the most. They are each amazing ladies, and they inspire me to strive for my goals in spite of obstacles. I am honored to be their older sister and lifelong friend.

Finally, my deepest appreciation is extended to my parents, Everett and Jewell Sharpe. I would like to thank them for their eternal love and support. Through their lives, they have shown me the importance of discipline, perseverance, and hard work, which have been essential throughout this process. Their encouragement of my endeavors throughout my life has motivated me to pursue each of my ambitions with excellence and confidence. I am forever grateful to my parents, and I could not have completed this thesis without them.

TABLE OF CONTENTS

page

ACKNOWLEDGMENTS 4

LIST OF TABLES 7

LIST OF FIGURES 8

LIST OF ABBREVIATIONS 9

ABSTRACT 10

CHAPTER

1 BACKGROUND 12

Introduction 12
Current Influenza Surveillance in the United States 13
Demographics of Internet and Social Media Users 16
Google Data for Influenza Surveillance 18
Twitter Data for Influenza Surveillance 21
Wikipedia Data for Influenza Surveillance 23
Objectives 25

2 METHODS 26

Data Collection 26
Statistical Analysis 27

3 RESULTS 30

Sample Characteristics 30
Summary of Change Points Detected in the 2012-2015 Influenza Seasons 31
Comparison of Change Points Detected among Web-based Sources 32
Comparison of Sensitivity and Positive Predictive Value among Web-based Sources 32

4 DISCUSSION 38

Key Findings 38
Limitations 39
Implications and Future Research 40
Conclusions 41

LIST OF REFERENCES 43

BIOGRAPHICAL SKETCH 50

LIST OF TABLES

Table page

3-1 Summary of weekly influenza-like illness (ILI) count data for the CDC, Google, Twitter, and Wikipedia, 2012-2015 influenza seasons 34

3-2 Comparison of change points detected using Bayesian change point analysis, 2012-2015 influenza seasons 35

3-3 Comparison of sensitivity and positive predictive value (PPV) among web-based sources, 2012-2015 influenza seasons 36

LIST OF FIGURES

Figure page

3-1 Change points detected by Bayesian change point analysis, 2012-2013 influenza season..... 36

3-2 Change points detected by Bayesian change point analysis, 2013-2014 influenza season..... 37

3-3 Change points detected by Bayesian change point analysis, 2014-2015 influenza season..... 37

LIST OF ABBREVIATIONS

API	application programming interface
CDC	Centers for Disease Control and Prevention
ILI	influenza-like illness
ILINet	United States Outpatient Influenza-like Illness Surveillance Network
MMWR	Morbidity and Mortality Weekly Report
PPV	positive predictive value

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Science

SURFING FOR FLU: COMPARING GOOGLE, TWITTER, AND WIKIPEDIA AS TOOLS FOR INFLUENZA SURVEILLANCE USING BAYESIAN CHANGE POINT ANALYSIS

By

Joyce Danielle Sharpe

May 2016

Chair: Robert L. Cook
Major: Epidemiology

Research has shown some promise of using data from Google, Twitter, and Wikipedia to complement traditional surveillance for influenza-like illness (ILI). However, past studies have evaluated these web-based sources without comparing all three of them. The objective was to comparatively analyze Google, Twitter, and Wikipedia by examining which best corresponds with CDC ILI data. It was hypothesized that Wikipedia will best correspond with CDC data since prior research found it is the least influenced by media reporting.

Publicly available, de-identified data were collected from the CDC, Google Flu Trends, HealthTweets, and Wikipedia for the 2012-2015 influenza seasons. Bayesian change point analysis was used to detect seasonal changes in each of the data sources. Detections in Google, Twitter, and Wikipedia were compared to detections in CDC data, which was the gold standard. All analyses were conducted using the R package ‘bcp’ v4.0.0 in RStudio v0.99.484.

During the 2012-2015 influenza seasons, 40% of Wikipedia’s change points aligned with changes detected in the CDC data. 85% of Google’s change points were

identified concurrently with seasonal changes in the CDC data. 43% of Twitter's change points coincided with change points found in CDC data.

Google had the most changes that matched the CDC, disproving my hypothesis. Findings demonstrated that Google, Twitter, and Wikipedia occasionally correspond well with seasonal changes captured in CDC data, yet these sources are not reliable for detecting all seasonal changes in CDC data and should be further studied and developed.

CHAPTER 1
BACKGROUND

Introduction

Influenza is a contagious, viral respiratory illness that is spread person-to-person through respiratory droplets [1]. It is recognized by the following symptoms: fever, cough, sore throat, runny or stuffy nose, muscle or body aches, headaches, fatigue, and, less commonly, vomiting and diarrhea [2]. The influenza season in the United States (U.S.) is generally considered to last from October through May, peaking between December and February [3]. Although vaccine-preventable, influenza is a costly and, often, severe burden on the U.S. health care system, causing 3,000-50,000 deaths annually [4-5]. The most vulnerable populations for influenza infections are the elderly, pregnant women, young children, and people with compromised immune systems [1]. As one of many influenza surveillance systems, the Centers for Disease Control and Prevention (CDC) monitors influenza cases using syndromes of influenza-like illness (ILI) that are reported by sentinel healthcare providers comprised of physicians, physician assistants, and nursing professionals.

The CDC defines ILI as a fever (≥100°F/37.8°C) and a cough and/or sore throat without a known cause other than influenza [6]. Since ILI is indistinguishable from influenza without laboratory testing, most data reported by the CDC are reported as ILI incidence, not influenza incidence [7]. Since traditional influenza surveillance relies on ILI syndromes that are reported by healthcare providers, this sort of surveillance method primarily captures those who readily seek medical care for their influenza symptoms, thus missing those who do not interact with the health care system. In addition, traditional surveillance methods are plagued by dated technology and reporting delays

of up to one to two weeks [8; Sophie Smith, email communication, December 2015]. Syndromic surveillance, which can be defined as the monitoring of disease syndromes in or near real-time for early detection of outbreaks, has incorporated the use of novel data sources, such as emergency department records and prescription sales, to enhance traditional surveillance systems [9-11]. However, with increased numbers of people using the Internet, it is necessary to further study other novel, digital surveillance approaches to complement traditional methods.

With over three billion Internet users around the world [12], data mining web-based sources can provide valuable epidemic intelligence by identifying mentions and searches of various illnesses that could represent the people that are missed by traditional systems. Recently, nontraditional data sources, particularly those that are Internet-based and open-source, have come into greater application for public health surveillance. This is especially evident as individuals who experience various symptoms may search online for health-related information and share their illness experiences using social media platforms before seeking medical care. Using such data sources has been coined as digital epidemiology. Digital epidemiology is known as the use of web-based sources, such as social media sites, online news reports, and search query records, for public health surveillance [13-15]. Digital epidemiology methods can be less expensive and more timely as well as expand the range of health events that can be detected for public health surveillance.

Current Influenza Surveillance in the United States

Influenza surveillance in the U.S. is a cooperative endeavor between the Epidemiology and Prevention Branch in the Influenza Division at the CDC and partnering healthcare providers, clinics, emergency departments, laboratories, and

state, local, and territorial health departments [6]. Data for influenza surveillance is collected and evaluated continuously throughout the year, assembled in a weekly report known as FluView. Influenza surveillance is necessary in order to monitor influenza-related illnesses, detect when and where influenza activity is happening nationally, and understand the impact of influenza on hospitalizations and mortality [6]. Also, surveillance is needed to distinguish what strains of the influenza virus are circulating and identify changes in influenza viruses in order to better develop vaccines [6]. The CDC compiles data on five categories of surveillance to analyze influenza activity: virologic surveillance, geographic spread of influenza, hospitalization surveillance, mortality surveillance, and outpatient illness surveillance [6].

There are nearly 350 clinical and public health laboratories that contribute to virologic surveillance for influenza [6]. Participating laboratories report the weekly total of respiratory specimens tested as well as the quantity and percent of specimens that tested positive for influenza [6]. After becoming a nationally notifiable condition in 2007, special attention has been placed on the detection of novel influenza A viruses due to the pandemic potential of these virus strains [6]. Moreover, state health departments monitor the weekly level of geographic spread of influenza illness for their respective states using the State and Territorial Epidemiologists Report [6]. Geographic spread is reported as one of the following levels: no activity, sporadic, local, regional, or widespread [6].

The Influenza Hospitalization Surveillance Network tracks laboratory-confirmed influenza-related hospitalizations in children and adults by reviewing hospital documents and databases [6]. Also, mortality surveillance is conducted by monitoring influenza-

associated deaths from death certificate data, and also pediatric influenza and pneumonia deaths. The National Center for Health Statistics and the122 Cities Mortality Reporting System both report influenza-related mortality data from state vital statistics offices and vital statistics offices of 122 U.S. cities, respectively [6]. Specifically, these systems report weekly death certificate data and the proportion of deaths due to pneumonia or influenza, which are compared to a seasonal baseline and epidemic threshold [6]. Established as a nationally notifiable condition in 2004, influenza-associated pediatric deaths that are confirmed positive for influenza by laboratory testing are reported through the mortality surveillance system as well [6].

For outpatient illness surveillance, the U.S. Outpatient Influenza-like Illness Surveillance Network (ILINet) is used to monitor patient visits to healthcare providers for influenza activity [6]. Nearly 2,000 participating outpatient healthcare providers report the total weekly number of patient visits and the proportion of patients seen for ILI [6]. To prepare a national estimate, the proportion of patient visits for ILI is weighted by state population to account for differences in the variability of state population and provider contribution by state (Sophie Smith, email communication, December 2015). Specifically, weighted ILI rates are calculated by the following equations (Sophie Smith, email communication, December 2015). Firstly, unweighted, state-specific ILI rates are calculated. For example,

*Florida (FL) %ILI = (sum of ILI visits reported by FL providers / sum of patient visits reported by FL providers) * 100*

*Georgia (GA) %ILI = (sum of ILI visits reported by GA providers / sum of patient visits reported by GA providers) * 100*

Then, the unweighted, state-specific ILI rates are multiplied by the respective state population based on U.S. Census data.

$$FL\ \%ILI * FL\ population = X_1$$

$$GA\ \%ILI * GA\ population = X_2$$

Lastly, this value is calculated for each state, summed together, and then divided by the total population. The resulting value is the weighted ILI rate.

$$(X_1 + X_2 + \ldots + X_n) / total\ population = weighted\ national\ ILI\ rate$$

To determine if ILI activity is higher than normal, weighted ILI rates are compared each week to the national baseline, which is calculated by the CDC. The baseline is determined by computing the average proportion of patient ILI visits during non-influenza weeks for the previous three influenza seasons plus two standard deviations [6]. The CDC suggests against applying the national baseline level to regional data due to variations in regional data, thus specific regional baselines are calculated using the same methodology for the national baseline [6]. Furthermore, as an online application managed by the CDC, FluView Interactive provides free data exploration and downloading privileges for data from each of the five components of the U.S. influenza surveillance system [16].

Demographics of Internet and Social Media Users

Understanding the demographics of Internet users in the U.S. is important in order to fully comprehend the opportunities and limitations of data collected from web-based sources. Internet use is widespread in America, increasing from 14% of the population to nearly 90% in a 20-year period [17]. In 2014, it was reported that a

majority (87%) of American adults are Internet users [17]. Internet use in the U.S. does not vary much based on sex, race/ethnicity, and living environment. 86% of American women use the Internet, while 87% of men in the U.S. are web users [17]. Regarding race and ethnicity, 81% of African Americans use the Internet compared to 83% of Hispanics and 85% of Caucasians [17]. Also, according to the Pew Research Center, Americans living in rural settings use the Internet at similar levels (83%) to Americans living in urban areas (88%) and suburban locales (87%) [17].

There are some disparities in Internet use when evaluating age groups, education levels, and household incomes. 97% of Americans between the ages of 18-29 are Internet users, which is comparable to 93% of those who are ages 30-49 and 88% of people ages 50-64 [17]. However, only 57% of Americans age 65 years and older use the Internet [17]. Only 76% of Americans with a high school education or less demonstrate Internet usage as opposed to 97% of those with at least a college education and 91% of those with some college education [17]. According to the Pew Research Center, Americans with household incomes of at least $50,000 a year exhibit high Internet use (93%-99%) compared to Americans with household incomes of less than $30,000 a year (77%) [17].

The proportion of Americans who access the Internet using a mobile device has increased from 29% of cell phone owners in 2009 to 60% of cell phone owners in 2013 [18]. In addition, Americans under the age of 18 years old use the Internet at high rates [19]. 95% of teenagers in the U.S. have access to the Internet, and 74% of teens have access to the web using a mobile device [19]. On the contrary, 14% of Americans do not use the Internet [20]. According to the Pew Research Center, a majority of non-

Internet users are African American (20%) or Hispanic (17%), 65 years of age and older (41%), have not completed high school (44%), live in rural areas (21%), and earn less than $30,000 a year in household income (25%) [20].

Furthermore, since Twitter has been studied as a potential tool for influenza surveillance, it is important to evaluate the demographics of its user population. Twitter saw an increase in adult users from 18% in 2013 to 23% in 2015 [21]. Nearly one in five female Internet users subscribe to Twitter, while one in four online men use Twitter [21]. About 6% of Internet users aged 65 years and older use Twitter compared to over 30% of Internet users aged 18-29 [21]. Duggan also concluded that Twitter is more popular among Americans with a college degree (27%) when compared to those with a high school education or less (19%) [21]. According to Duggan, 28% of African American Internet users subscribe to Twitter, which is similar to 28% of Hispanic Internet users [21]. Only 20% of Caucasian Internet users subscribe Twitter [21]. 30% of Internet users who subscribe to Twitter live in urban areas compared to 15% of rural Internet users [21]. Overall, increased use of the Internet and social media platforms can be beneficial for public health surveillance as more health-related information is searched for and shared on the web.

Google Data for Influenza Surveillance

Globally, millions of people conduct daily online searches for health-related information, making search query data valuable resources of public health data for surveillance purposes [22]. The first known instance of using a digital data source for public health surveillance was when Dr. Gunther Eysenbach conducted analyses of health-related searches on the Internet [23-24]. Through his work, Eysenbach introduced the concept of infodemiology, or information epidemiology, as the study of

the distribution and determinants of health information on the Internet, such as Google searches, in order to inform public health [14]. In 2006, Eysenbach found a strong correlation (Pearson's r=0.91) between Google search trend data and the 2004-2005 influenza season in Canada [25]. In 2008, Google established its Flu Trends application to monitor frequency changes in online searches related to influenza [13; 26-27].

Google Flu Trends is a system that data mines de-identified, aggregate, and proprietary Google search data for mentions of influenza-like symptoms to generate real-time estimates of influenza incidence [26; 28-29]. Search data were collected using an automated method that identified relevant searches using 45 query keywords [26]. Google Flu Trends also collects data for dengue, a mosquito-borne infectious disease [29-30]. Overall, Google Flu Trends estimates have shown correlations with official influenza surveillance data from the CDC, while consistently delivering estimates one to two weeks ahead of CDC reports [13; 26]. Although initially successful in earlier years, the system has not been without its issues, more recently. For instance, during the 2012-2013 influenza season, Google Flu Trends overestimated influenza incidence at a rate double that of CDC surveillance reports, and the system significantly underestimated influenza incidence during the 2009 H1N1 influenza pandemic [28; 31-33]. Due to issues with its proprietary algorithm, Google Flu Trends was discontinued in August 2015 [29].

Moreover, studies have investigated the effectiveness of Google Flu Trends at varying geographic levels. In one study, Google search query data and CDC weekly ILI data were compared at the city (New York City, New York), regional (Mid-Atlantic), and national (U.S.) scale [33]. Olson et al. found that both the original (2008) and revised

(2009) algorithms for Google Flu Trends were not generally reliable at each geographic level, particularly in instances of varying intensity in influenza seasons and media coverage [33]. However, another study reported that local-level Twitter data best predicted confirmed influenza cases in an inner-city hospital (Baltimore, Maryland), while national-level Google Flu Trends data performed best at detecting ILI cases at any geographic scale [34]. Overall, the Broniatowski et al. study found that when combined, Google Flu Trends and Twitter data are most effective at detecting overall influenza activity at the national scale [34].

Also, one study investigated the practicality of using Google Flu Trends data to develop an early warning system. Pervaiz et al. conducted a comparative analysis of three different epidemic detection algorithms – normal distribution, Poisson distribution, and negative binomial distribution algorithms – on detecting influenza outbreaks using data from Google Flu Trends [27]. They compared the sensitivity, specificity, timeliness, and alerting accuracy of each algorithm, finding that there was no single best method [27]. Normal distribution methods, which are based on historical limits and cumulative sums, have been traditionally used for influenza surveillance by the CDC [27]. However, Pervaiz et al. found that Poisson- and negative binomial-based models are more effective than normal distribution algorithms for mitigating the impact of noise on the quality of search query data, especially with variances in Internet users and their activities [27].

In addition, a study explored the link between Google Flu Trends and prescription sales data, finding a strong correlation (Pearson’s r=0.92) between prescription sales data and Google Flu Trends and an even stronger correlation (Pearson’s r=0.97)

between prescription sales data and CDC ILI data [35]. Google search query data is not limited to studying influenza, but can also be used for studying mental health [36], rotavirus [37], and purchasing behavior of cigarettes [38]. As a widely used source for digital epidemiologic analyses, Google search query data must be evaluated to encompass the impact that changes in influenza seasonality, Internet demographics, and online search behaviors will have when developing future algorithms to estimate influenza incidence [27-28; 31-33]. All in all, with the ability to detect health trends in digital spaces, Google search query data has presented an innovative and cost-effective approach to complement traditional influenza surveillance methods.

Twitter Data for Influenza Surveillance

Founded in 2006, Twitter is a social media platform that allows its users to freely create and share tweets, or messages that are limited to 140 characters [39-40]. As of the second quarter of 2015, Twitter reported an average of 304 million active users [41]. As of 2013, active Twitter users have posted over 170 billion tweets [40]. To access tweets for analysis or research, Twitter provides an application programming interface (API) that allows for the search and retrieval of public Twitter data only [39-40; 42]. Public tweets are those that are accessible to anyone, while protected tweets are private and only available to a protected user's Twitter network [43]. A 1% random sample of all public tweets are available via the API [44]. Twitter also provides a geolocation feature for tweets. About 2-3% of all tweets are geotagged, while only 3% of users register for the geolocation feature [39-40; 44].

Influenza-related tweets have been shown to be highly correlated with reported ILI incidence [45]. A study described that when compared to sentinel physician ILI rates and emergency department ILI rates, tweets that mentioned "flu" or "influenza" were

closely correlated to the general trend for both data sources [46]. However, it was found that emergency department ILI data had a stronger correlation to tweets than ILI data from sentinel medical providers did [46]. Another study showed that tweets better reflect the incidence of influenza in comparison to search queries. Aramaki and colleagues compared a Twitter-based epidemic detection system to a similar system based on Google Flu Trends [47]. Aramaki et al. discovered that during periods of high news coverage, both methods were vulnerable to excessive media coverage, being less correlated to influenza incidence in Japan [47]. However, when compared to influenza data in Japan, Twitter data (Pearson's r=0.89) outperformed data from Google Flu Trends (Pearson's r=0.847) during periods of normal news coverage [47].

Also, studies have found that keyword choice influences how tweets compare to CDC ILI data. Natural language processing (NLP) is a field of artificial intelligence that programs computers to understand human (natural) language. Allowing for a deeper content analysis of tweets, the use of NLP has demonstrated the importance of keyword choice when conducting analysis of tweets [47-49]. By using NLP, Nagel et al. reported that tweets mentioning "flu" were better correlated with CDC ILI rates than tweets mentioning "influenza" [49]. One study contradicted the importance of keyword choice by finding that even seemingly irrelevant tweets about influenza have a moderate correlation with CDC ILI rates [7]. However, primarily using NLP to classify tweets presents some challenges because machine learning techniques are not as effective as human interpreters in understanding slang, regional language, foreign language, misspellings, vernacular, current events, and other aspects of language that require contextual knowledge [50].

Moreover, geographic limitations of Twitter data present difficulties for using the social networking site for influenza surveillance. Abstracting and evaluating tweets at minute geographic levels can be difficult as too few tweets are actually geotagged, especially at the municipal or county level [51]. One study, however, has demonstrated the ability of tweets to have a trend with city-level ILI data, but that study used Twitter data from New York City, a major metropolitan area with an abundance of Twitter users who can generate ample tweets to be used for trend analysis [52]. In spite of these drawbacks, still, studies on using Twitter for influenza surveillance have generally found correlation between influenza-related tweets and national CDC ILI data [46-47; 52].

Wikipedia Data for Influenza Surveillance

Established in 2001, Wikipedia is an online, user-regulated encyclopedia that contains a vast quantity of freely available informative articles. These 30 million encyclopedia-like articles are created, modified, and maintained by Internet users in any of Wikipedia's 287 languages [8; 53-56]. The vastest and most used edition of Wikipedia is the English language edition with nearly 4.9 million articles [8; 56]. Wikipedia is highly trafficked, reporting about 31,000 total active editors and over 500 million visitors each month [54]. Wikipedia allows hourly page view data that dates back to 2007 to be freely accessible via Wikimedia Statistics [53-54; 57-58].

Wikipedia page view data has proven valuable for tracking trending topics and the development of breaking news stories as well as disease monitoring and forecasting [53; 57]. A study has compared view data of influenza-related Wikipedia articles to official CDC influenza activity reports, finding that Wikipedia's page view data accurately estimates national influenza incidence up to two weeks before the CDC within a reasonable margin of error [54]. McIver and Brownstein demonstrated that Wikipedia's

page view data on influenza-related articles are more accurate at weekly estimates than Google Flu Trends and that it performs well during abnormal influenza seasons and high media reporting, which are issues for Google Flu Trends and, to a lesser extent, Twitter [54]. During peak activity of ILI incidence, Wikipedia traffic, and media reporting, a model of 35 influenza-related Wikipedia articles was highly correlated with CDC ILI data (Pearson's $r=0.946$, $p<0.001$), while the model was also highly correlated with CDC ILI data (Pearson's $r=0.958$, $p<0.001$) in the absence of such activity [54]. One study found that Wikipedia page view data has suitable forecasting value up until the peak of the influenza seasons [57]. Another study also reported that Wikipedia page view data is suitable for forecasting using a 28-day analysis as well as for nowcasting, or monitoring current disease incidence [53].

However, there can be disadvantages with using Wikipedia data for public health surveillance. Since Wikipedia has become a preferred source for seeking health information [59-60], the signal-to-noise ratio of Wikipedia data can be problematic [53]. Generous et al. explained that although some users search Wikipedia in order to understand possible influenza illness, other users search Wikipedia from informative purposes or purposes not related to influenza infection, misconstruing the influenza incidence that can be captured by Wikipedia [53]. Also, Wikipedia's hourly updating feature can be too incongruent for the surveillance of certain diseases, depending on the disease's progression [53]. For example, Generous and colleagues found that since tuberculosis and HIV/AIDS progress slowly as infections, the hourly quality of Wikipedia data does not capture the intensity of these diseases well [53]. Another issue that researchers can encounter with using Wikipedia is that its page view data may include

“multiple views on the same article by the same user” [54]. Although extensive research has not been conducted on this novel data source, studies by McIver and Brownstein, Hickmann et al., and Generous et al. have showed that Wikipedia has some potential for public health surveillance [53-54; 57].

Objectives

Studies have shown some promise of using data from Google, Twitter, and Wikipedia to complement traditional surveillance systems in order to understand the epidemiology of influenza. However, past research has evaluated these web-based sources individually without any comparison amongst all three of them. This study seeks to conduct a comparative analysis of Google, Twitter, and Wikipedia by examining which digital data source is most aligned with ILI data from the CDC ILINet system. Specifically, Bayesian change point analysis is used to identify change detections in each data source in order to evaluate how comparable data from Google Flu Trends, Twitter, and Wikipedia are to CDC ILI data. For this study, the specific research question is the following: for which web-based source – Google Flu Trends, Twitter, or Wikipedia – do detected change points most closely match change points detected in national-level CDC ILI data for the 2012-2013, 2013-2014, and 2014-2015 influenza seasons? My hypothesis is that Wikipedia data will have the most change points in common with CDC ILI data due to research by McIver and Brownstein [54] finding that Wikipedia data can be less influenced by media coverage in comparison to data from Google Flu Trends and Twitter, to a lesser degree [33; 47].

CHAPTER 2
METHODS

Data Collection

Data were retrospectively collected for the 2012-2013 influenza season (September 30, 2012 – May 18, 2013), 2013-2014 influenza season (September 29, 2013 – May 17, 2014), and 2014-2015 influenza season (September 28, 2014 – May 23, 2015) [61-63]. All data were presented as MMWR weeks. MMWR (Morbidity and Mortality Weekly Report) weeks represent the various weeks of the epidemiologic year for reporting disease incidence [64]. MMWR weeks start on Sunday and end on Saturday, ranging from 1 to 53 weeks with most years encompassing 52 MMWR weeks [64]. Each of the influenza seasons included in this study begins in MMWR week 40 and ends in MMWR week 20 of the following year.

CDC ILINet data were freely downloaded from FluView Interactive. ILINet data are de-identified and aggregated by MMWR weeks. The CDC ILINet system collects information on counts of patients seen for ILI, totals of all patients seen by participating healthcare providers, number of providers, ILI counts by age group, unweighted and weighted ILI rates, and corresponding year and week [16]. National-level counts of ILI patient visits were used for this study. For Google, de-identified, national-level count data of influenza-related Google searches made in the U.S. were freely downloaded from the Google Flu Trends website [29]. Data from Google Flu Trends were already aggregated by MMWR weeks.

For data from Twitter, de-identified, national-level count data of influenza-related Twitter conversations in the U.S. were freely downloaded from HealthTweets.org, a Johns Hopkins University-based repository of influenza-related tweets dating back to

November 2011 [65]. Using the Twitter API, HealthTweets data are collected from a keyword stream, which is 1% of public tweets [65]. Dredze et al. [65] categorized influenza-related tweets using automated annotators that are based on keywords, keyword combinations, and the classifier developed by Lamb and colleagues [48]. A random 1% of all public tweets are also collected to normalize the influenza-related tweets gathered from the keyword stream [65]. On the HealthTweets site, registered users can view trends of tweets in the U.S. and United Kingdom [65]. Trends can even been further differentiated by U.S. state [65]. Data from HealthTweets were also already aggregated by MMWR weeks.

Wikipedia article view data, which are de-identified and aggregated, were gathered for the “Influenza” article (English version). National-level Wikipedia data that was collected was restricted to the U.S. Wikipedia has made its article view data freely available for downloading through Wikimedia Statistics [58]. Wikipedia data is presented as the number of article views by the hour, which includes non-unique views [58]. Since Wikipedia article view data on the “Influenza” article were only aggregated by the hour in Wikimedia Statistics, the data were aggregated by MMWR weeks before analysis.

Statistical Analysis

For this study, Bayesian change point analysis was the method used to detect change points for each of the data sources for the 2012-2015 influenza seasons. All Bayesian change point analyses were conducted using the R package ‘bcp’ version 4.0.0 [66-68] in RStudio version 0.99.484 (RStudio Inc., Boston, MA) [69]. This Bayesian method provides the probability of a change point at any given interval (i.e. MMWR week) in time series data [70]. The probability of a change in time series data is

found using a complex Markov Chain Monte Carlo (MCMC) procedure [66-68; Dr. John Emerson, email communication, March 2016].

Specifically, the ‘bcp’ package is an MCMC implementation of the Bayesian change point method described by Barry and Hartigan [70], providing estimates of the posterior mean at each interval as well as the probability of a change at each given interval [66-68]. In Bayesian analysis, the posterior mean is the mean of the posterior distributions of a given interval in a time series. Posterior distributions are computed from the Bayes’ rule [66; 70], which is defined by the following equation [71]:

$$P(A|B) = [P(B|A)\ P(A)] \div P(B),$$

where P(A) = 0.2. P(A) represents the prior distribution of the time series. The default prior distribution used in Bayesian change point analysis is 0.2 because this hyperparameter (p_0) was found to be suitable for various types of time series processes [66-67; 70].

As the default, the ‘bcp’ package conducts 500 MCMC iterations after excluding 50 burn-in iterations to compute the posterior mean and posterior probability of a change point [66-68]. MCMC iterations are used in the estimation of the posterior mean for any given interval [66-68]. Burn-in iterations are the initial iterations of a Markov chain implementation, which are removed to diminish the effect of initial iterations on the posterior mean and probability [66-68]. The probability of a change point at any given interval in a time series is computed from the number of times in the MCMC iterations that the time series met the condition of having a change point at that interval [Dr. John Emerson, email communication, March 2016; Dr. Chandra Erdman, email communication, March 2016].

Change points were defined as detections where the probability of a change point was greater than or equal to 50%. Change points detected in the CDC ILI data were used as the gold standard for comparison of change point detections. Detections of the web-based sources that occurred during the exact week, one preceding week, or one week after the CDC detections were considered matching, or true, change points. This was done to account for any reporting lags, which can be common with surveillance data. The number of change points for each data source was compared, and sensitivity and positive predictive values were calculated for the web-based sources.

CHAPTER 3
RESULTS

Sample Characteristics

A summary of count data collected from the CDC, Google Flu Trends, Twitter, and Wikipedia for the 2012-2015 influenza seasons is shown in Table 3-1. The 2014-2015 influenza season consisted of 34 total MMWR weeks because MMWR week 53 occurred during the season due a preceding calendar leap year. The 2012-2013 and 2013-2014 influenza seasons both consisted of 33 MMWR weeks and included data up to MMWR week 52, which is common for most epidemiologic years since most calendar years have 52 weeks in total.

The data collected for each of the data sources showed some variability. The CDC data showed less variation over the three influenza seasons compared to the web-based sources. In the 2012-2013 influenza season, the CDC averaged 19,049 ILI counts per week, while averaging 16,574 weekly ILI counts and nearly 20,000 weekly ILI counts during the 2013-2014 and 2014-2015 influenza seasons, respectively. Google count data had a higher weekly count average in the 2012-2013 season (4,121 searches per week) than in the 2013-2014 and 2014-2015 influenza seasons, which averaged 2,274 weekly searches and 2,549 weekly searches, respectively.

For Twitter, the average number of weekly ILI tweets decreased with each influenza season. In the 2012-2013 influenza season, the average number of ILI tweets was 8,096 per week, whereas in the 2013-2014 influenza season, the average fell to 5,826 tweets per week and then declined to 2,900 tweets per week during the 2014-2015 season. Wikipedia also saw similar variability. The site averaged 47,541 page views per week during the 2012-2013 influenza season. However, Wikipedia's average

dropped to 25,039 weekly page views and nearly 22,000 page views per week in the 2013-2014 and 2014-2015 influenza seasons, respectively. See Table 1 for further data summary information.

Summary of Change Points Detected in the 2012-2015 Influenza Seasons

For the 2012-2013 influenza season, 14 total change points were identified by Bayesian change point analysis (Figure 3-1). Google and the CDC both had the least number of detections with three change points each. Change points were found for the CDC in MMWR weeks 48, 50, and 5 and Google in MMWR weeks 51, 4, and 5. Twitter and Wikipedia both had four detected change points in MMWR weeks 47, 1, 3, and 4 and MMWR weeks 52, 1, 3, and 5, respectively.

Bayesian change point analysis found 17 change points amongst the four data sources for the 2013-2014 influenza season (Figure 3-2). Google and Twitter had the most number of change points with five each. Changes in Google data were detected in MMWR weeks 48, 50, 51, 3, and 5, while Twitter had change points detected for MMWR weeks 48, 51, 4, 7, and 17. Four change points were detected for the CDC (MMWR weeks 48, 50, 6, and 15), and three change points were found for Wikipedia (MMWR weeks 51, 1, and 6).

In the 2014-2015 influenza season, Bayesian change point analysis detected a total of 18 change points (Figure 3-3). Five change points were detected for the CDC in MMWR weeks 48, 49, 50, 53, and 6. Five change points were also detected for both Google (MMWR weeks 48, 50, 51, 53, and 4) and Twitter (MMWR weeks 43, 50, 2, 3, and 12). The Bayesian analysis only identified three change points for Wikipedia, which were in MMWR weeks 44, 53, and 3.

Comparison of Change Points Detected among Web-based Sources

A comparison of change points detected in each of the data sources is provided in Table 3-2. For the 2012-2013 influenza season, Google had three total change points in common with the CDC, which were MMWR weeks 51, 4, and 5. Twitter had two change points in common with the CDC's detections, which were MMWR weeks 47 and 4. Wikipedia had only one detection that matched a CDC change point, which was MMWR week 5.

In the 2013-2014 influenza season, Google had a total of four change points (MMWR weeks 48, 50, 51, and 5) that coincided with changes detected in the CDC data. Twitter had three detections (MMWR weeks 48, 51, and 7) that matched change points in the CDC data. Wikipedia had two change points in common with CDC ILI data, which were MMWR weeks 51 and 6.

For the 2014-2015 influenza season, four change points (MMWR weeks 48, 50, 51, and 53) were detected in the Google data that concurred with changes identified in the CDC data. Both Twitter and Wikipedia had only one change point detection that coincided with the change points found in the CDC data, which were MMWR weeks 50 and 53, respectively.

Comparison of Sensitivity and Positive Predictive Value among Web-based Sources

A comparison of sensitivity and positive predictive values (PPV) among the web-based sources is provided in Table 3-3. Sensitivity and PPV were computed using the change points detected for web-based sources that matched change points detected for the CDC (true positives), change points detected by the CDC but not by the web-based sources (false negatives), and change points detected by the web-based sources but

not by the CDC (false positives). Specifically, sensitivity was calculated by dividing the true positives for each web-based source by the total of true positives and false negatives, which would be the total number of CDC change points [72]. PPV was calculated by dividing the true positives for each web-based source by the total of true positives and false positives, which would be the total change points for that particular web-based source [72]. A high sensitivity of 92% was found for Google, while the positive predictive value for Google was 85%. A low sensitivity of 50% was calculated for Twitter; a low PPV of 43% was found for Twitter also. Wikipedia had a low positive predictive value of 40% and a low sensitivity of 33%.

Table 3-1. Summary of weekly influenza-like illness (ILI) count data for the CDC, Google, Twitter, and Wikipedia, 2012-2015 influenza seasons

	CDC	Google	Twitter	Wikipedia
2012-2013 Influenza Season				
MMWR Weeks	33	33	33	33
Mean (counts/week)	19049	4121	8096	47541
Min (counts/week)	7317	1286	2558	29865
Max (counts/week)	39896	10555	22935	114919
2013-2014 Influenza Season				
MMWR Weeks	33	33	33	33
Mean (counts/week)	16574	2274	5826	25039
Min (counts/week)	9033	1339	1196	17885
Max (counts/week)	28654	5008	10506	36935
2014-2015 Influenza Season*				
MMWR Weeks	34	34	34	34
Mean (counts/week)	19940	2549	2900	21918
Min (counts/week)	9289	1144	451	12958
Max (counts/week)	40664	6911	8709	35232

*While most epidemiologic years are comprised of 52 MMWR weeks, the 2014-2015 influenza season included MMWR week 53, which is why it has a total of 34 MMWR weeks and not 33 MMWR weeks.

Table 3-2. Comparison of change points detected using Bayesian change point analysis, 2012-2015 influenza seasons

Influenza Season	CDC ILI counts (ref)	Google counts	Twitter counts	Wikipedia counts
2012-2013			Week 47*	
	Week 48			
	Week 50			
		Week 51*		
				Week 52
			Week 1	Week 1
			Week 3	Week 3
		Week 4*	Week 4*	
	Week 5	Week 5*		Week 5*
2013-2014	Week 48	Week 48*	Week 48*	
	Week 50	Week 50*		
		Week 51*	Week 51*	Week 51*
				Week 1
		Week 3		
			Week 4	
		Week 5*		
	Week 6			Week 6*
			Week 7*	
	Week 15			
			Week 17	
2014-2015			Week 43	
				Week 44
	Week 48	Week 48*		
	Week 49			
	Week 50	Week 50*	Week 50*	
		Week 51*		
	Week 53	Week 53*		Week 53*
			Week 2	
			Week 3	Week 3
		Week 4		
	Week 6			
			Week 12	

*Indicates a corresponding change point to the CDC change points (reference).

Table 3-3. Comparison of sensitivity and positive predictive value (PPV) among web-based sources, 2012-2015 influenza seasons

	Sensitivity*	Positive predictive value**
Google	92%	85%
Twitter	50%	43%
Wikipedia	33%	40%

*Sensitivity was calculated as: true change points for a source / total change points for the gold standard.
**PPV was calculated as: true change points for a source / total change points for a source.

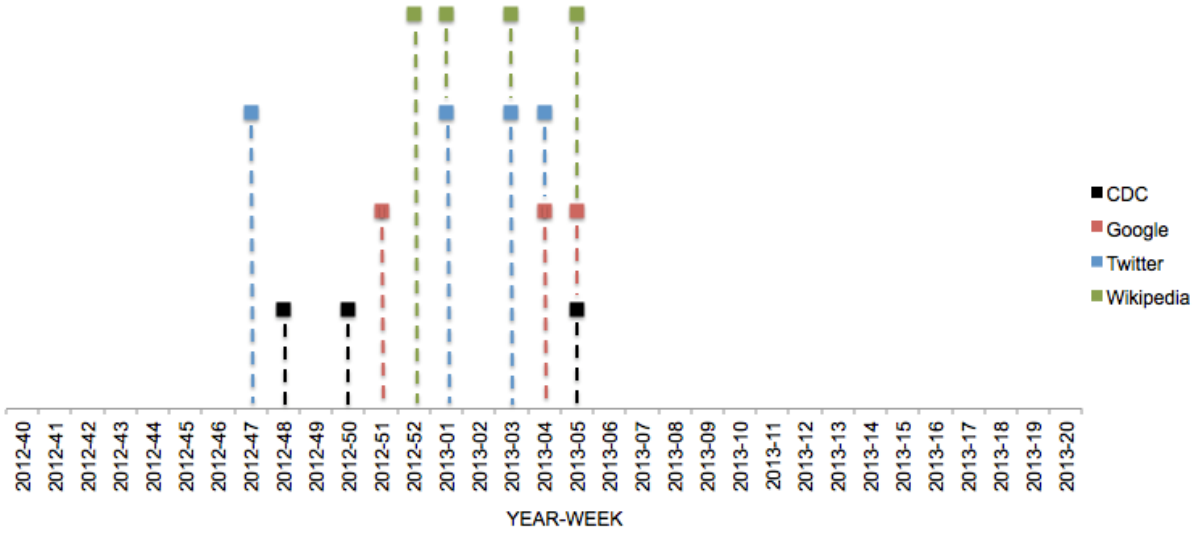

Figure 3-1. Change points detected by Bayesian change point analysis, 2012-2013 influenza season.

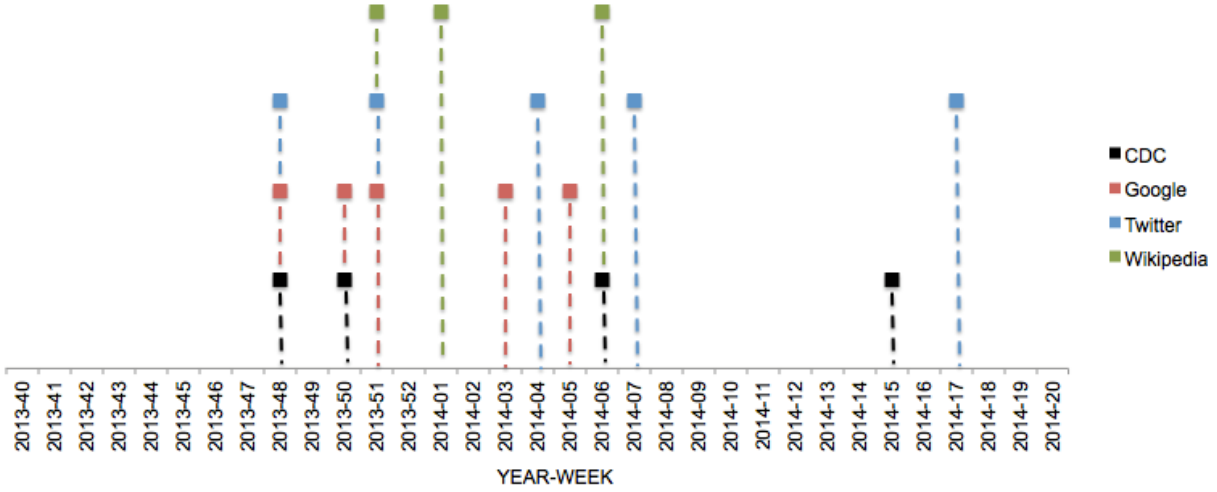

Figure 3-2. Change points detected by Bayesian change point analysis, 2013-2014 influenza season.

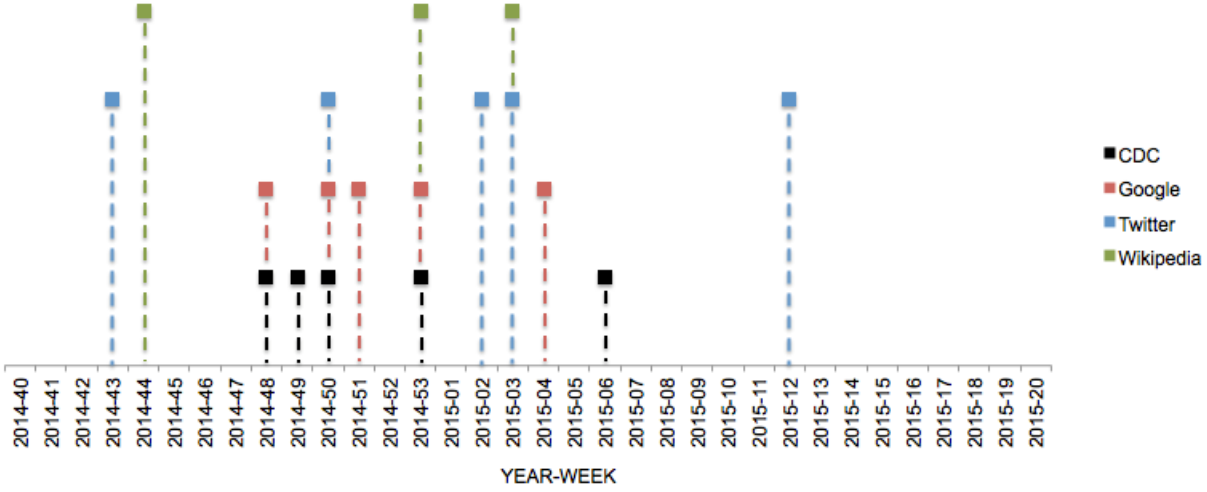

Figure 3-3. Change points detected by Bayesian change point analysis, 2014-2015 influenza season.

CHAPTER 4
DISCUSSION

Key Findings

Google had a total of 11 true change points (three in the 2012-2013 influenza season, four in the 2013-2014 influenza season, and four in the 2014-2015 influenza season) that coincided with the CDC's change points. Twitter had six matching change points with the CDC data, and Wikipedia had only four matching change points. Since Google had the most change points that coincided with changes detected in the CDC ILI data, my hypothesis that Wikipedia would have the most change detections was not supported.

Furthermore, sensitivity and positive predictive value are important for evaluating the quality of surveillance data [72]. Calculating sensitivity and positive predictive values for each web-based source revealed the variability amongst the sources. Google had a moderate positive predictive value and was highly sensitive, while Twitter and Wikipedia both had low sensitivity rates and positive predictive values.

Evaluating the detected change points demonstrates that Google, Twitter, and Wikipedia have some detections that align well with CDC ILI data; however, they do not identify all changes that were identified in the CDC data, which would be important for understanding when seasonal changes occur during an influenza season. Also, from assessing sensitivity and positive predictive values for the web-based sources, it is clear that web-based sources need to be further studied and developed before being incorporated as surveillance data to complement a traditional system.

Limitations

There are limitations in this study that are necessary to address. Firstly, Bayesian change point analysis assumes normality about time series data [73]. Since public health surveillance data can be variable and have non-normal distribution, the normal distribution assumption can be problematic when estimating the posterior distribution [73]. Another major limitation to using Bayesian change point analysis is that it cannot be used as a technique to analyze real-time data [74]. Bayesian change point analysis is best used to evaluate changes in historical time series data after all data have been collected. For this study, the Bayesian method was used to retrospectively evaluate data collected from the CDC, Google, Twitter, and Wikipedia after each influenza season occurred; therefore, the results cannot be directly applied to influenza surveillance in reality where data is collected and analyzed prospectively.

In addition, there may be some disadvantages to the data used in this study. CDC ILI data can have its own limitations. Typically each week, only about 1,800 of over 2,900 participating ILINet providers report ILI data to the CDC [6; 54]. This can be problematic in regards to data completeness, but the CDC attempts to account for this by weighting ILI rates by the number of reporting providers and state population [6]. Also, the algorithms developed to extract searches and tweets for Google Flu Trends and HealthTweets still require improvements [33; 48], which could have affected the quality of the count data that was collected from both of these sources.

Moreover, only the Wikipedia “Influenza” article was used for analysis, excluding other articles on influenza medications and influenza strains. A study has evaluated the utility of other influenza-related Wikipedia articles for surveillance purposes [54], but those were not included in this study. Also, Wikipedia article view data include non-

unique views, thus one user can have multiple views that contribute to high counts in Wikipedia's data [58]. However, both Google Flu Trends and HealthTweets do not distinguish or remove multiple searches and tweets that can be credited to a single user [26; 65]. The non-unique nature of the web-based sources may not be problematic if each non-unique instance represents a separate experience of ILI syndromes, but this cannot be confirmed with existing research and should be further investigated.

Lastly, the demographics of Internet users are not comparable to the demographics of the U.S. population. Internet users are, on average, younger than the general American population [17]. Although this difference may be viewed as a limitation to using web-based data for public health surveillance, younger age groups (0-4 years, 5-24 years, and 25-49 years) are overrepresented in outpatient ILI counts that are reported to the CDC [16]. However, older populations (50-64 years and >=65 years) are overrepresented in ILI-related hospitalization rates that are reported to the CDC [16], which are not captured in web-based data.

Implications and Future Research

The findings in this study show that changes in data from web-based sources may not be completely parallel with changes in CDC ILI data throughout an influenza season. Although not evaluated in real-time or for prospective use, the web-based sources have demonstrated that they do well with detecting changes during the weeks (MMWR weeks 47-51) leading up to the peaks (MMWR week 52 or 53, depending on the season year) of the influenza seasons included in this study. This has practical implications because if web-based sources can detect the peak of an influenza season before official CDC data, then public health practice could monitor data from Google, Twitter, and Wikipedia to identify a season's peak. Data from web-based sources can

also inform practicing health professionals when to expect a height in outpatient ILI visits, which could further direct when to increase influenza vaccine production and immunization promotion campaigns.

Furthermore, there is more substantive information in the content of web-based data that is not accounted for in count data. Future research has already begun to conduct content analyses of web-based data in order to understand what Internet users experience with various illnesses. Content analyses have proven valuable for both communicable and non-communicable diseases because Internet users share and search about health experiences from mental health [75-76] to substance use [77-78] to health needs of sexual minorities [79]. Therefore, future studies should emphasize conducting content analyses of posts, blogs, and other communication on the Internet.

Conclusions

This study contributes to the scientific literature because, to the author's knowledge, it is the first study to comparatively evaluate Google, Twitter, and Wikipedia as possible data sources for influenza surveillance. Previous studies have compared Twitter to Google [47; 50] and Wikipedia to Google [54] in regards to correlating with CDC ILI data, and there has even been a study combining web-based sources in prediction models to forecast influenza activity in the U.S. [80]. However, this is the first study to evaluate and compare the surveillance utility of all three web-based sources – Google, Twitter, and Wikipedia.

Moreover, it is significant to assess Google, Twitter, and Wikipedia as tools for influenza surveillance because previous research has studied each of them individually or only done a dual comparison. Yet, comparing Google, Twitter, and Wikipedia can help to inform public health professionals if these web-based sources accurately reflect

the seasonal changes in CDC ILI data and whether or not web-based data sources should be used to complement traditional systems for influenza surveillance. In closing, web-based data sources occasionally have detections that correspond well with seasonal changes captured in CDC ILI data, but, currently, such data sources are not reliable in detecting all seasonal changes found in CDC ILI data and should be further studied and developed.

LIST OF REFERENCES

1. Seasonal influenza: flu basics. Centers for Disease Control and Prevention. 2015. Available: http://www.cdc.gov/flu/about/disease/index.htm.

2. Flu symptoms & severity: influenza symptoms. Centers for Disease Control and Prevention. 2014. Available: http://www.cdc.gov/flu/about/disease/symptoms.htm.

3. The flu season. Centers for Disease Control and Prevention. 2014. Available: http://www.cdc.gov/flu/about/season/flu-season.htm.

4. Chunara R, Goldstein E, Patterson-Lomba O, Brownstein JS. Estimating influenza attack rates in the United States using a participatory cohort. Sci Rep. 2015;5:9540. doi: 1038.srep09540

5. Centers for Disease Control and Prevention. Estimates of deaths associated with seasonal influenza – United States, 1976-2007. MMWR Morb Mortal Wkly Rep. 2010;59(33):1057-1062.

6. Overview of influenza surveillance in the United States. Centers for Disease Control and Prevention. 2015. Available: http://www.cdc.gov/flu/weekly/overview.htm.

7. Bodnar T, Salathé M. Validating models for disease detection using Twitter. In: Proceedings of the 22nd International Conference on World Wide Web Companion. 2013;699-702.

8. Fairchild G, De Silva L, Del Valle SY, Segre A. Eliciting disease data from Wikipedia articles. In: 9th International AAAI Conference on Web and Social Media. 2015;1-33.

9. Buehler JW, Hopkins RS, Overhage JM, Sosin DM, Tong V. Framework for evaluating public health surveillance systems for early detection of outbreaks: recommendations from the CDC Working Group. MMWR Recomm Rep. 2004;53(RR05):1-11.

10. Buehler JW, Sonricker A, Paladini M, Soper P, Mostashari F. Syndromic surveillance practice in the United States: findings from a survey of state, territorial, and selected local health departments. Adv Dis Surveill. 2008;6(3):1-20.

11. Henning KJ. Overview of syndromic surveillance: what is syndromic surveillance? MMWR Suppl. 2004;53(Suppl):5-11.

12. ICT facts and figures: the world in 2015. International Telecommunication Union. 2015. Available: http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf.

13. Bernardo TM, Rajic A, Young I, Robiadek K, Pham MT, Funk JA. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. J Med Internet Res. 2013;15(7):e147. doi: 10.2196/jmir.2740

14. Eysenbach G. Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. J Med Internet Res. 2009;11(1):e11. doi: 10.2196/jmir.1157

15. Salathé M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, et al. Digital epidemiology. PLoS Comput Biol. 2012;8(7):e1002616. doi: 10.1371/journal.pcbi.1002616

16. FluView Interactive. Centers for Disease Control and Prevention. 2016. Available: http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html.

17. Internet user demographics. Pew Research Center. 2014. Available: http://www.pewinternet.org/data-trend/internet-use/latest-stats/.

18. Cell phone activities. Pew Research Center. 2013. Available: http://www.pewinternet.org/data-trend/mobile/cell-phone-activities/.

19. Internet user demographics: teen Internet access demographics. Pew Research Center. 2012. Available: http://www.pewinternet.org/data-trend/teens/internet-user-demographics/.

20. Offline adults. Pew Research Center. 2013. Available: http://www.pewinternet.org/data-trend/internet-use/offline-adults/.

21. Duggan M. The demographics of social media users. Pew Research Center. 2015. Available: http://www.pewinternet.org/2015/08/19/the-demographics-of-social-media-users/.

22. Johnson HA, Wagner NM, Hogan WR, Chapman W, Olszewski RT, Dowling J, et al. Analysis of web access logs for surveillance of influenza. *Stud Health Technol Inform*. 2004;107(pt. 2):1202-1206.

23. Eysenbach G, Kohler C. What is the prevalence of health-related searches on the World Wide Web? Qualitative and quantitative analysis of search engine queries on the Internet. In: AMIA Annual Symposium Proceedings. 2003;225-229.

24. Eysenbach G, Kohler C. Health-related searches on the Internet. JAMA. 2004;291(24):2946-2946. doi: 10.1001/jama.291.24.2946

25. Eysenbach G. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. In: AMIA Annual Symposium Proceedings. 2006;244-248.

26. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009;457(7232):1012-1014. doi: 10.1038/nature07634

27. Pervaiz F, Pervaiz M, Abdur Rehman N, Saif U. FluBreaks: Early epidemic detection from Google Flu Trends. J Med Internet Res. 2012;14(5):e125. doi: 10.2196/jmir.2102

28. Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google Flu Trends performance in the United States during the 2009 influenza virus A (H1N1) pandemic. PLoS ONE. 2011;6(8):e23610. doi: 10.1371/journal.pone.0023610

29. Google Flu Trends. Google, Inc. n.d. Available: http://www.google.org/flutrends/about/.

30. Gluskin RT, Johansson MA, Santillana M, Brownstein JS. Evaluation of Internet-based dengue query data: Google Dengue Trends. PLoS Negl Trop Dis. 2014;8(2):e2713. doi: 10.1371/journal.pntd.0002713

31. Butler D. When Google got flu wrong. Nature. 2013;494(7436):155-156. doi: 10.1038/494155a

32. Lazer D, Kennedy R, King G, Vespignani A. The parable of Google Flu: traps in big data analysis. Science. 2014;343(6176):1203-1205. doi: 10.1126/science.1248506

33. Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L. Reassessing Google Flu Trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. PLoS Comput Biol. 2013;9(10):e1003256. doi: 10.1371/journal.pcbi.1003256

34. Broniatowski DA, Dredze M, Paul MJ, Dugas A. Using social media to perform local influenza surveillance in an inner-city hospital: a retrospective observational study. JMIR Public Health Surveillance. 2015;1(1):e5. doi: 10.2196/publichealth.4472

35. Patwardhan A, Bilkovski R. Comparison: Flu prescription sales data from a retail pharmacy in the US with Google Flu Trends and US ILINet (CDC) data as flu activity indicator. PLoS ONE. 2012;7(8):e43611. doi: 10.1371/journal.pone.0043611

36. Ayers JW, Althouse BM, Allem JP, Rosenquist JN, Ford DE. Seasonality in seeking mental health information on Google. Am J Prev Med. 2013;44(5):520-525. doi: 10.1016/j.amepre.2013.01.012

37. Desai R, Lopman BA, Shimshoni Y, Harris JP, Patel MM, Parashar UD. Use of Internet search data to monitor impact of rotavirus vaccination in the United States. Clin. Infect. Dis. 2012;54(9):e115-e118. doi: 10.1093/cid/cis121

38. Ayers JW, Ribisl K, Brownstein JS. Using search query surveillance to monitor tax avoidance and smoking cessation following the United States' 2009 “SCHIP” cigarette tax increase. PLoS ONE. 2011;6(3):e16777. doi: 10.1371/journal.pone.0016777

39. Burton SH, Tanner KW, Giraud-Carrier CG, West JH, Barnes MD. "Right time, right place" health communication on Twitter: value and accuracy of location information. J Med Internet Res. 2012;14(6):e156. doi: 10.2196/jmir.2121

40. Leetaru K, Wang S, Padmanabhan A, Shook E. Mapping the global Twitter heartbeat: the geography of Twitter. First Monday. 2013;18(5). doi: 10.5210/fm.v18i5.4366

41. Twitter: Number of monthly active users 2010-2015. Statista. 2015. Available: http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.

42. The streaming APIs. Twitter, Inc. 2015. Available: https://dev.twitter.com/streaming/overview.

43. About public and protected tweets. Twitter, Inc. n.d. Available: https://support.twitter.com/articles/14016.

44. Dredze M, Paul MJ, Bergsma S, Tran H. Carmen: a Twitter geolocation system with applications to public health. In: Proceedings of the AAAI Workshop on Expanding the Boundaries of Health Informatics Using Artificial Intelligence. 2013.

45. Broniatowski DA, Paul MJ, Dredze M. Twitter: big data opportunities. Science. 2014;345(6193):148.

46. Aslam AA, Tsou M-H, Spitzberg BH, An L, Gawron JM, Gupta DK, et al. The reliability of tweets as a supplementary method of seasonal influenza surveillance. J Med Internet Res. 2014;16(11):e250. doi: 10.2196/jmir.3532

47. Aramaki, E., Maskawa, S., & Morita, M. Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the 2011 Conference on Empirical Natural Language Processing Conference (EMNLP). 2011;1568-1576.

48. Lamb, A., Paul, M. J., & Dredze, M. Separating fact from fear: tracking flu infections on Twitter. In: Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013;789-795.

49. Nagel AC, Tsou M-H, Spitzberg BH, An L, Gawron JM, Gupta DK, et al. The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. J Med Internet Res. 2013;15(10):e237. doi: 10.2196/jmir.2705

50. Nagar R, Yuan Q, Freifeld CC, Santillana M, Nojima A, Chunara R, et al. A case study of the New York City 2012-2013 influenza season with daily geocoded Twitter data from temporal and spatiotemporal perspectives. J Med Internet Res. 2014;16(10):e236. doi: 10.2196/jmir.3416

51. Paul MJ, Dredze M. You are what you tweet: Analyzing Twitter for public health. In: Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM). 2011.

52. Broniatowski DA, Paul MJ, Dredze M. National and local influenza surveillance through Twitter: an analysis of the 2012-2013 influenza epidemic. PLoS ONE. 2013;8(12):e83672. doi: 10.1371/journal.pone.0083672

53. Generous N, Fairchild G, Deshpande A, Del Valle SY, Priedhorsky R. Global disease monitoring and forecasting with Wikipedia. PLoS Comput Biol. 2014;10(11):e1003892. doi: 10.1371/journal.pcbi.1003892

54. McIver D, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol. 2014;10(4):e1003581. doi: 10.1371/journal.pcbi.1003581

55. Wikipedia. Wikimedia Foundation. 2015. Available: https://en.wikipedia.org/w/index.php?title=Wikipedia&oldid=636552708.

56. Wikipedia statistics. Wikimedia Foundation. 2015. Available: http://stats.wikimedia.org/EN/Sitemap.htm.

57. Hickmann KS, Fairchild G, Priedhorsky R, Generous N, Hyman JM, Deshpande A, et al. Forecasting the 2013–2014 influenza season using Wikipedia. PLoS Comput Biol. 2015;11(5):e1004239. doi: 10.1371/journal.pcbi.1004239

58. Page view statistics for Wikimedia projects. Wikimedia Foundation. 2015. Available: http://dumps.wikimedia.org/other/pagecounts-raw/

59. Laurent MR, Vickers TJ. Seeking health information online: does Wikipedia matter? J Am Med Inform Assoc. 2009;16(4):471-479. doi: 10.1197/jamia.M3059

60. Tausczik Y, Faasse K, Pennebaker JW, Petrie KJ. Public anxiety and information seeking following the H1N1 outbreak: blogs, newspaper articles, and Wikipedia visits. Health Commun. 2012;27(2):179-185. doi: 10.1080/10410236.2011.571759

61. Centers for Disease Control and Prevention. Influenza Activity – United States, 2012-2013 Season and Composition of the 2012-14 Influenza Vaccine. MMWR Morb Mortal Wkly Rep. 2013;62(23):473-479.

62. Epperson S, Blanton L, Kniss K, Mustaquim D, Steffens C, Wallis T, et al. Influenza Activity – United States, 2013-2014 Seasons and Composition of the 2014-2015 Influenza Vaccines. MMWR Morb Mortal Wkly Rep. 2014;63(22):483-490.

63. Appiah GD, Blanton L, D'Mello T, Kniss K, Smith S, Mustaquim D, et al. Influenza activity – United States, 2014-2015 season and composition of the 2015-2016 influenza vaccine. MMWR Morb Mortal Wkly Rep. 2015;64(21):583-590.

64. National notifiable diseases surveillance system: MMWR weeks. Centers for Disease Control and Prevention. n.d. Available: http://wwwn.cdc.gov/nndss/document/MMWR_Week_overview.pdf.

65. Dredze M, Cheng R, Paul MJ, Broniatowski DA. HealthTweets.org: a platform for public health surveillance using Twitter. In: Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.

66. Erdman C, Emerson JW. bcp: an R Package for performing a Bayesian analysis of change point problems. J Stat Softw. 2007;23(3):1-13.

67. Erdman C, Emerson JW. A fast Bayesian change point analysis for the segmentation of microarray data. Bioinformatics. 2008;24(19):2143-2148.

68. Wang X, & Emerson, JW. Bayesian change point analysis of linear models on general graphs, working paper. 2015.

69. RStudio Team. RStudio: integrated development for R. RStudio, Inc., Boston, MA. 2015. Available: http://www.rstudio.com/.

70. Barry D, Hartigan JA. A Bayesian analysis for change point problems. JASA. 1993;88(421):309-319.

71. Stone JV. Bayes' rule: a tutorial introduction to Bayesian analysis. 1st ed. Sheffield: Sebtel Press; 2013.

72. German RR, Lee LM, Horan JM, Milstein RL, Pertowski CA, Waller MN, et al. Updated guidelines for evaluating public health surveillance systems: recommendations from the Guidelines Working Group. MMWR Recomm Rep. 2001;50(RR13):1-35.

73. Kass-Hout TA, Xu Z, McMurray P, Park S, Buckeridge DL, Brownstein JS, et al. Application of change point analysis to daily influenza-like illness emergency department visits. J Am Med Inform Assoc. 2012;19(6):1075-1081. doi: 10.1136/amiajnl-2011-000793.

74. Kass-Hout TA. Change point analysis: introduction. Google Sites. 2013. Available: https://sites.google.com/site/changepointanalysis/home.

75. Cavazos-Rehg PA, Krauss MJ, Sowles S, Connolly S, Rosas C, Bharadwaj M, et al. A content analysis of depression-related tweets. Comput Human Behav. 2016;54:351-357.

76. Reavley NJ, Pilkington PD. Use of Twitter to monitor attitudes toward depression and schizophrenia: an exploratory study. Abdullah J, ed. PeerJ. 2014;2:e647. doi:10.7717/peerj.647

77. Chen AT, Zhu S-H, Conway M. What online communities can tell us about electronic cigarettes and hookah use: a study using text mining and visualization techniques. Eysenbach G, ed. J Med Internet Res. 2015;17(9):e220. doi: 10.2196/jmir.4517.

78. Krauss MJ, Sowles SJ, Moreno M, Zewdie K, Grucza RA, Bierut LJ, et al. Hookah-related Twitter chatter: a content analysis. Prev Chronic Dis. 2015;12:e121. doi: 10.5888/pcd12.150140.

79. Krueger EA, Young SD. Twitter: A Novel Tool for Studying the Health and Social Needs of Transgender Communities. Eysenbach G, ed. JMIR Mental Health. 2015;2(2):e16. doi: 10.2196/mental.4113

80. Santillana M, Nguyen AT, Dredze M, Paul MJ, Nsoesie EO, Brownstein JS. Combining search, social media, and traditional data sources to improve influenza surveillance. PLoS Comput Biol. 2015;11(10):e1004513. doi: 10.1371/journal.pcbi.1004513

BIOGRAPHICAL SKETCH

Joyce Danielle Sharpe was born in Savannah, Georgia. She was raised in three southern towns in Georgia: Baxley, Brunswick, and Waycross. Joyce Danielle graduated in 2011 from Ware County High School in Waycross, Georgia as the first African American valedictorian in the school's history. She completed her undergraduate education as a President's Scholar at the Georgia Institute of Technology, graduating in the spring of 2014 with a Bachelor of Science with Highest Honors in history, technology, and society with a concentration in the history of medicine. Joyce Danielle received her Master of Science in epidemiology from the University of Florida in the spring of 2016. Throughout her graduate studies at the University of Florida, Joyce Danielle upheld academic excellence, was involved with researching mobile health avenues to assist people living with HIV, and had early experiences in the professoriate, teaching Critical Thinking in Health Care, and Ethical and Legal Issues in Public Health and Health Professions. After graduating from the University of Florida, Joyce Danielle will be attending Emory University to pursue a Ph.D. in epidemiology starting in the fall of 2016.