ABSTRACT

Title of Document: GOING VIRAL: INTERNET AND SOCIAL MEDIA BASED SURVEILLANCE SYSTEMS FOR DETECTING INFLUENZA ACTIVITY IN MARYLAND

Lisa Marie Bowen, MPH Epidemiology, 2015

Directed By: Professor and Chair, Dr. Robert S. Gold, Department of Epidemiology and Biostatistics

Influenza surveillance is essential for detecting and managing outbreaks. The Maryland Department of Health and Mental Hygiene (DHMH) currently includes the number of emergency room and physician visits for influenza-like-illness (ILI) to track flu activity. Recently, internet and social media based surveillance methods have emerged as useful in detecting outbreaks. This study aims to determine if internet and social media based surveillance methods are useful in monitoring ILI in Maryland through assessing how Google Flu Trends (GFT) and tweets compare to portions of DHMH's formal reporting system. Innovations of this study include using symptom based keywords and incorporating a variety of sources of surveillance data. Results show tweets had a strong positive correlation with all other surveillance sources, Pearson's correlation coefficients ranged from 0.62-0.68. GFT were more highly correlated with DHMH data. Further research should investigate automating collection of tweets, application to other diseases, and standardized methods for location determination.

GOING VIRAL: INTERNET AND SOCIAL MEDIA BASED SURVEILLANCE SYSTEMS FOR DETECTING INFLUENZA ACTIVITY IN MARYLAND

By

Lisa Marie Bowen.

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Masters of Public Health
2015

Advisory Committee:
Professor Dr. Robert S. Gold, Chair
Dr. Sandra C. Quinn
Dr. Xin He

Dedication

This thesis is dedicated to Blaine, who was (almost) always willing to listen to my progress and frustrations, provided encouragement, and was nice enough to resist the temptation of watching our television shows without me.

Acknowledgements

Research reported in this paper was supported by Leidos Research Fellowship Program. The content is solely the responsibility of the author and does not necessarily represent the official views of the Leidos Corporation. Thank you to the Leidos team members who provided guidance for my research. I would also like to thank my committee members for their assistance and mentoring during my thesis process. I appreciate the efforts of David Blythe, Anikah Salim, and Andrea Bankoski at the Maryland Department of Health and Mental Hygiene in helping me obtain influenza surveillance data.

Table of Contents

.

List of Tables

List of Figures

Chapter 1: Introduction to Influenza Surveillance

The influenza pandemic of 1918 killed conservatively 21 million people worldwide, more people than the black plague, and the majority of deaths occurred within 24 weeks. Rapid mutation and antigen shift of influenza makes novel strains of the virus a continuous threat. Since 1918, six other pandemic influenzas have emerged, although none as lethal as the "Spanish flu" (1).

In 2005, the Federal Government developed a strategy for pandemic influenza. This strategy stresses the importance of real time (at onset of illness) surveillance in detecting and efficiently managing outbreaks (2). Currently, the Centers for Disease Control and Prevention (CDC) and the Maryland Department of Health and Mental Hygiene (DHMH) provide weekly reports on a variety of clinical data, including the number of emergency room visits due to influenza-like-illness (ILI). Recently, other forms of surveillance, such as Google search queries have emerged as useful in detecting outbreaks. Google Flu Trends detection shows increases in ILI symptoms 1-2 weeks ahead of ILI surveillance reports by the CDC (3). Many studies on disease surveillance mention the advantage of using multiple sources of surveillance to enhance effectiveness in early detection of outbreaks (4–8). For instance, at the onset of the H_1N_1 outbreak, informal internet based surveillance systems were reporting events before health organizations (9).

Social media provides another form of internet surveillance to track outbreaks (5–7,10,11). Social media supplies unique information for disease surveillance apart from formal reporting and Google searches by providing access to real time data from

individuals themselves, who may not be seeking medical care, or searching for online health information related to their symptoms.

Specific Aims

The long term goal of this project is to improve preparedness for influenza pandemics by using surveillance techniques that provide the earliest detection of outbreaks. This is an exploratory study that will illustrate challenges, priorities, and strategies associated with utilizing Twitter for ILI surveillance and contribute to the growing body of research on using social media for disease surveillance. Twitter is a social media platform where users share messages, called tweets, which are a maximum of 140 characters in length. The terms Twitter data, tweets, and Twitter messages will be used interchangeably throughout the manuscript. This study goes beyond measuring an association, instead the purpose of this study is to explore new datasets that were not known a decade ago in order to investigate new strategies to improve upon and strengthen standard practice in the field of epidemiologic surveillance. The objective of this study is to determine if internet and social media surveillance methods are useful in monitoring ILI in Maryland. The objective is further divided into three specific aims. **Aim 1**: Determine similarity between influenza-like-illness emergency department and physician visits for 2014-2015 flu season to past flu seasons in Maryland. This will show how comparable the 2014-2015 flu season is to other flu seasons. Since only one season of Twitter data will be used in this study, this aim will provide evidence for the correlation between DHMH and Twitter data in a typical flu season. **Aim 2**: Assess how Google Flu Trend data for influenza in Maryland compares to portions of DHMH's formal reporting system.

This aim will investigate whether or not Google Flu Trend data are useful tool for detecting ILI in Maryland. **Aim 3**: Examine Twitter (a widely used social media source) messages to determine if they could be used as a source of influenza surveillance data by assessing the correlation with DHMH and Google data and determine if they provide more timely information on influenza outbreaks. **Subaim 3.1:** Analyze the characteristics of Twitter users to investigate whether or not certain sub-portions of the population are under or over represented. **Subaim 3.2:** Explore characteristics of Tweets to determine how correlation varies and compare Twitter data to DHMH data on laboratory confirmed cases to gain a better insight on the variety of ways Twitter data can be used as a surveillance tool. Since traditional surveillance is limited to people seeking health care, internet based surveillance methods provide a way to strengthen current systems by overcoming this limitation (3,7,11). For instance, if the majority of people who have the flu self-treat at home; traditional surveillance methods will miss the majority of cases. In addition, a retrospective study on the use of social media and internet surveillance methods in tracking the 2010 Haiti cholera outbreak found informal sources were highly correlated with official data, but provided more immediate access to information due to delays in obtaining official reports (5). A multi-faceted approach to influenza surveillance has the potential to improve and provide more rapid response to outbreaks (6–9). This study is important because it aims to determine if internet based surveillance methods (Google Flu Trends and Twitter) are useful in monitoring influenza-like-illness in Maryland. Favorable results of this study have important implications for emergency preparedness and planning procedures.

Literature Review

Transmission of Influenza through droplets that can infect people up to six feet away means pandemic causing strains can spread quickly, especially in the interconnected world we live in today (12). This makes early detection vital to saving the most lives by limiting outbreaks and identifying the causative strain to manufacture vaccines.

Many studies have begun to mention the importance of social media and internet surveillance in tracking outbreaks (3,5–8,13). Specifically, using social media can provide information on early outbreaks, as well as monitor public concerns (7). Some social media such as Twitter is also easier to use by researchers and professionals due to the proprietary nature of Google (10). Twitter was used by a Chicago health department during food borne illness outbreaks to link possible cases to an internet reporting form. Subsequently, researchers found the majority of potential cases who filled out forms did not seek medical treatment, and would not have been included if only traditional surveillance methods had been used (11).

All studies that have evaluated social media and internet surveillance have found a correlation with CDC data and a more immediate detection of outbreaks (3,5,13,14). Corley et al. searched all internet blogs for keywords related to influenza, and when compared to CDC influenza-like-illness data, researchers calculated a Pearson's correlation coefficient of $r = 0.63$ (13). Ginsberg et al.'s comparison of Google Flu Trends and CDC influenza-like-illness data resulted in a very high ($r = 0.90$) correlation (3). Achrekar and colleagues assessed mentions of influenza related keywords on Twitter and calculated a correlation of $r = 0.98$ with CDC data

(14). An investigation of Twitter content during the H_1N_1 pandemic found data from Twitter predicted outbreaks 1-2 weeks ahead of the CDC on average (10). A report by Ginsberg et al. found Google Flu Trends was also ahead of the CDC by 1-2 weeks in terms of estimating weekly influenza activity (3).

A variety of limitations in using these informal surveillance methods have also been revealed. Schmidt pointed out that surveillance relying on Google search queries may be susceptible to noise, like graduate students researching the flu, decreasing its reliability as a method to detect outbreaks (10). While Google has been shown to be highly correlated with CDC data for seasonal influenza, it was found to have low correlation with formal data during the onset of the 2009 H_1N_1 pandemic (7). One challenge in using social media to track outbreaks is that social media contains a large number of news reports, instead of "self identified" influenza information (6). Twitter is more popular among young, college educated people. Therefore, analyses using Twitter data have the potential to over represent these groups and under represent other sub-groups such as minorities and the elderly (15). In addition, the correlation between Twitter data and confirmed influenza cases has yet to be established (6,10,13). Another limitation in using Twitter is location estimation from users, only approximately 1% of tweets contain geo-coded location information (16,17). Therefore, other information should be used to determine the location of Twitter users. While no standardized method for location estimation exists, previous studies have concluded that time zone information is more reliable than location entries in determining the location of a user/tweet (17,18).

All studies using Twitter data have focused on keywords associated with a certain influenza strain or words “influenza” and “flu”. A recent keyword search of Tweets using “influenza” and “flu” found a multitude of Tweets related to flu vaccines and news/information. This relates to the challenge mentioned by Salathe et al. and Corley et al. that many Tweets don’t contain “self-identified” influenza information (6,13). An innovation of this proposed study is using key words consistent with the influenza-like-illness case definition, fever (cough OR sore throat) (19). Using this combination of keywords should provide more data on “self identified” illness and help eliminate Tweets on general flu information. A study in 2010 found no income or racial disparities in the general use of social networking sites, though strong disparities remained in internet access (20). More specifically, the PewResearch Internet Project shows a significant increase in Twitter usage among the 65+ population in 2014. In addition in 2014, 25% of online Hispanics and 27% of online African Americans used Twitter, compared to 21% of online whites (15). The increasing popularity of Twitter with a variety of demographic groups should reduce under-representation. However, an analysis of Twitter users will be included in this study to determine the demographic characteristics users included in this data set. Twitter data will also be compared to laboratory confirmed cases, a current gap in knowledge.

Chapter 2: Research Design and Methods

The purpose of this study was to investigate the association between different methods of influenza surveillance and to assess the usefulness of internet and social media based surveillance systems in monitoring influenza activity. This project was approved by the University of Maryland Institutional Review Board and was not considered human subjects research. All statistical analysis was performed using SAS software, version 9.3 of the SAS System for Windows (SAS Institute Inc., Cary, NC). Pearson's product moment correlation coefficient (referred to as Pearson's correlation coefficient through the remainder of the manuscript) was used to calculate the level of linear relationship of frequency per week reported by different surveillance methods. Pearson's correlation coefficient was chosen to enable comparisons as previous studies have set a precedent for using Pearson's correlation coefficient when analyzing Twitter data. Sample size was limited by the time period of official reporting of ILI symptoms (Oct- mid May, or more precisely Morbidity and Mortality Weekly Report (MMWR) weeks 40-20) and start of Twitter data collection (October 30, 2014). However, at least 20 weeks of data were collected for all surveillance sources. With 20 weeks of data and a type I error rate of 0.05, Pearson's correlation coefficients of 0.55 or higher can be detected with a power of 81.72%.

Data collection

Data were collected throughout the study as it became available from all sources. Data collection ended on March 28, 2015 due to project timeline requirements, since there is a one week delay in the release of influenza activity

reports from DHMH, the last week of DHMH data is for the week ending March 21, 2015. While the flu season does not officially end until the beginning of May, flu activity was considered minimal for seven consecutive weeks prior to the end of data collection according to DHMH weekly surveillance indicators (23).

Aim 1:

Maryland influenza-like-illness (ILI) surveillance data on emergency department visits, physician visits, and laboratory confirmed cases for the 2014-2015 flu season were obtained from the Maryland Weekly Influenza Surveillance Activity Reports available from the Maryland Department of Health and Mental Hygiene (DHMH) website. Weekly reports included activity from the previous week ("last week number") which usually differed from the activity level documented in the initial report ("this week number") due to delays in obtaining data. Figure 1 contains portions of the Weekly Influenza Surveillance Activity Reports from two consecutive weeks. Notice the columns marked with the arrows. The total ILI visits listed in "this week number" in the report for week ending March 7, 2015 corresponds to the number of total ILI visits in "last week number" for the report for week ending March 14, 2015. Since early detection is of primary interest in this study, in the event of a discrepancy between the numbers in the "this week" and "last week" columns, as in Figure 1, the number from the initial report ("this week number") was used. Last week numbers were used during weeks when no reports were released due to federal and state holidays.

Weekly Influenza Surveillance Activity Report week ending March 7, 2015

ILINet Sentinel Providers

Twenty-nine sentinel providers reported a total of 8,479 visits this week. Of those, 95 (1.1%) were visits for ILI. This is **below** the Maryland baseline of **2.0%**.

ILI Visits To Sentinel Providers By Age Group	This Week Number (%)	Last Week Number (%)	Season Number (%)
Age 0-4	27 (28%)	33 (20%)	1171 (25%)
Age 5-24	40 (42%)	67 (42%)	1939 (41%)
Age 25-49	18 (19%)	39 (24%)	1008 (21%)
Age 50-64	5 (5%)	17 (11%)	431 (9%)
Age ≥ 65	5 (5%)	5 (3%)	210 (4%)
Total ILI Visits	95 (100%)	161 (100%)	4759 (100%)

Weekly Influenza Surveillance Activity Report week ending March 14, 2015

ILINet Sentinel Providers

Twenty-seven sentinel providers reported a total of 10,714 visits this week. Of those, 132 (1.2%) were visits for ILI. This is **below** the Maryland baseline of **2.0%**.

ILI Visits To Sentinel Providers By Age Group	This Week Number (%)	Last Week Number (%)	Season Number (%)
Age 0-4	30 (23%)	27 (27%)	1211 (25%)
Age 5-24	44 (33%)	44 (44%)	1998 (41%)
Age 25-49	29 (22%)	19 (19%)	1040 (21%)
Age 50-64	17 (13%)	6 (6%)	449 (9%)
Age ≥ 65	12 (9%)	5 (5%)	222 (5%)
Total ILI Visits	132 (100%)	101 (100%)	4920 (100%)

Figure 1: Discrepancy in data from Weekly Influenza Activity Reports

Total number of positive rapid flu tests was used for laboratory confirmed cases. Data for number of positive rapid flu tests were from 32 clinical labs, rather than the DHMH lab administration, which resulted in a larger sample size (21). Influenza-like-illness surveillance data on emergency department visits, and physician visits from the 2014-2013, 2013-2012, 2012-2011, 2011-2010, 2010-2009, and 2009-2008 flu seasons were provided by DHMH from the Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE) system. DHMH data were combined into a Microsoft Excel file for statistical analysis.

Aim 2:

Google Flu Trend data have been approved for re-use and were downloaded from Google.org for use in this study. The downloaded dataset from Google.org had data from all states, and began in 2003. Only Google Flu Trend data from Maryland and from years that had corresponding DHMH data (2008-2015) were used.

Aim 3:

Tweets were collected from Twitter's Streaming API (Application Programming Interface) service via Tweetarchivist.com, a company offering subscriptions to provide publically available streaming Twitter data on specified keywords. The keyword combination "fever AND (cough OR sore throat)" was used to gather tweets related to influenza-like-illness. Data included characteristics such as username, location, time zone, date and time, and full Tweet text for each Tweet returned. The dataset of returned tweets was downloaded into a Microsoft Excel file

from Tweetarchivist.com four times throughout the data collection period resulting in four rounds of data cleaning as data became available to disperse the workload. A limit to Streaming API is not providing access to all of the Tweets related to the keywords. However, if the Tweets matching the keywords represent less than 1% of the total volume of Tweets, streaming API returns 100% of the matching Tweets (22). Since it is unlikely that the number of Tweets matching the keywords "fever (cough OR sore throat)" exceeded 1% of total Tweets, this was not a limitation of the current study.

Data cleaning

Aim 3:

All data cleaning was performed in Microsoft Excel (2010). Twitter data were cleaned to remove re-tweets, multiple tweets from one user in a 6-week time frame, and tweets occurring outside of the United States. Since incidence was of primary interest in this study, only original tweets were used. Re-tweets were identified and removed by searching for tweets containing "RT @" in the tweet text. Multiple tweets containing the same text from the same user were removed from the dataset as they were suspected to be bots (automated programmed posts) and not provide any information on an actual influenza case. If users had multiple original Tweets returned, Tweets were broken down into 6 week periods and only the first Tweet for each period was included in the final dataset. Six week periods were chosen based on how the CDC classifies new episodes of illness for surveillance reporting (14). This was done to help eliminate prevalence data and instead focus on the first incidence of

illness per user. If a user had a six-week time span between posts, then both posts were kept due to the ability to be re-infected with the influenza virus.

A previous study comparing Twitter Streaming API and Twitter firehose (full repository of Tweets) found the Streaming API returned a high percentage (90%) of geo-coded Tweets. However, geo-coded Tweets only represent a small minority of total Tweets, and can introduce bias (22,24). Therefore, data for the current study included tweets that were identified as occurring in the United States, not just Maryland. Time zone and location information were used to determine the location a tweet originated from. Previous studies have shown that time zone is more reliable than location in determining a user's location in the absence of geolocated data (17,18). Since only 0.97 percent of tweets returned were geolocated, location and time zone information were the main pieces of information used for location determination. The following rules were applied for determining which tweets most likely occurred in the United States, and therefore kept in the dataset. A flow chart containing the rules used for determining location can be found in Figure 2.

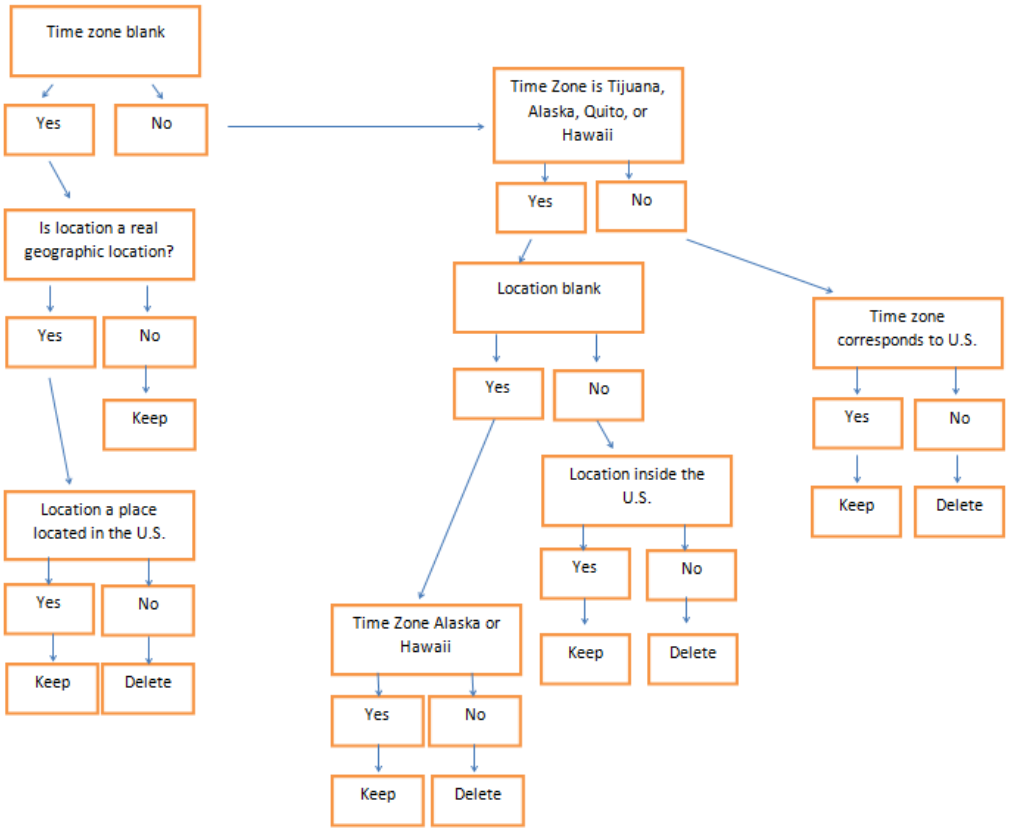

Figure 2: Flow chart for determining location of tweets

No standardized way to determine location from location and time zone data has been established. The rules used in this study were developed after reading existing literature and examining the data to create a standardized method to ensure the majority of tweets actually occurring in the United States were included in the dataset with minimal tweets from other countries being included (17,18). The process of location determination by hand was time-consuming and limits the application of Twitter data for use in public health settings unless automated procedures are developed. Therefore, the original data set underwent a separate data cleaning. For the second data cleaning method, re-tweets were removed and only one Tweet per user was included which reduced the amount of time needed to clean the data. The

correlation of frequency of tweets per day between the two data sets was then calculated using Pearson's correlation coefficient to determine if the data set which had minimal cleaning (referred to as the raw dataset) could be used as a proxy for tweets occurring in the United States.

After data cleaning, the remaining tweets were combined into one dataset, this dataset was then reformatted to allow comparisons to the other forms of surveillance data, which are recorded in frequency per week. Local time zone information was used to calculate the frequency of tweets per day. Frequency per day was then translated into frequency of tweets per week, based on MMWR weeks. A final dataset containing week ending date and frequency of tweets was then used in the analysis, see Table 1. This same formatting method was performed on the raw dataset in order to calculate the correlation coefficient between the two Twitter datasets.

Statistical Analysis

Emergency department visits, physician visits, and Google Flu Trend data were compared on a weekly basis from the first week of October until the end of May (MMWR week 40-20) for past flu seasons, 2008-2014, and until the week ending March 21 (MMWR week 11) for DHMH data and the last week in March (MMWR week 12) for Google Flu Trends and Twitter data for the 2014-2015 flu season. The final dataset used for analyzing the linear relationship between the different forms of influenza surveillance contained frequency per week for each surveillance method: tweets (raw and cleaned), Google Flu Trends, physician visits, emergency department visits, and laboratory confirmed cases (Table 1). A similar dataset containing frequency per week for physician visits, emergency department visits, and Google Flu

Trends from years 2008-2015 was used to calculate the correlation between the 2014-2015 flu season with past flu seasons. The same dataset was used in analysis of the linear relationship between Google Flu Trends and DHMH surveillance data (physician visits and emergency department visits) for each flu season dating back to the 2008-2009 flu season.

Table 1: Dataset of frequency per week for Aim 3 data analysis

Week Ending Date	Raw Twitter Data	Cleaned Twitter data	Google Flu Trends	Physician Visits	Emergency Department Visits	Laboratory Confirmed Cases
11/8/2014	939	623	2129	122	642	24
11/15/2014	913	607	1602	100	709	38
11/22/2014	879	559	1885	116	703	52
11/29/2014	835	552	2186	131	947	175
12/6/2014	942	650	2698	197	1114	301
12/13/2014	950	668	3340	254	1357	652
12/20/2014	987	706	4941	406	2265	2100
12/27/2014	1017	758	7536	293	3538	3307
1/3/2015	1030	717	8057	445	3394	2423
1/10/2015	989	689	6346	326	2298	1442
1/17/2015	860	594	5389	249	1494	920
1/24/2015	928	620	4358	282	1332	788
1/31/2015	951	654	4405	241	994	565
2/7/2015	913	600	3428	218	1028	514
2/14/2015	855	549	3139	167	926	312
2/21/2015	813	533	2516	132	771	258
2/28/2015	790	508	2228	159	723	203
3/7/2015	429	276	2099	95	620	136
3/14/2015	399	266	2125	132	744	161
3/21/2015	755	495	2134	120	802	183

Aim 1:

In order to determine if the 2014-2015 flu season was a typical flu season, emergency department visits and physician visits data from the 2014-2015 flu season were compared to each past flu season by calculating the Pearson's correlation coefficient which resulted in six different correlation coefficients. The correlation coefficients were then rank ordered.

Aim 2:

Pearson's correlation coefficient was also calculated to determine the correlation between Google Flu Trends and DHMH data. Since Google makes revisions to the algorithm used in Google Flu Trends, correlation coefficients were calculated for each flu season (3).

Aim 3:

Tweets were aggregated into frequency per week to be consistent with DHMH and Google Flu Trend's reporting methods. Twitter data was analyzed starting with MMWR week 45 (week ending 11/8/2014) as this was the first full week of Twitter data collected. Pearson's correlation coefficients were calculated for the correlation between Twitter data and emergency department visits, physician visits, Google Flu Trends, and laboratory confirmed cases, Table 1. Due to a lack of tweets with location and/or geo-coded information in Maryland, no separate analysis was performed comparing Maryland tweets to the full dataset.

Chapter 3: Results

The results demonstrate that internet and social media influenza surveillance methods are correlated with DHMH surveillance data on physician visits, emergency department visits, and laboratory confirmed cases. Results are further broken down and reported according to each aim of the study.

Aim 1

Aim 1 investigated the similarity between the 2014-2015 flu season to previous flu seasons. The objective of this aim was to determine if the linear relationship between tweets and DHMH data is generalizable to a typical flu season. The correlation coefficients between the 2014-2015 flu season and previous flu seasons varied dramatically; results are reported in ranked order by p-value according to physician visits in Table 2.

Table 2: Pearson's correlation coefficient between the 2014-2015 and past flu seasons

	Physician Visits	**Emergency Department Visits**
2012-2013	0.655 (p=0.0004)	0.820 (p<0.001)
2013-2014	0.485 (p=0.01)	0.546 (p=0.55)
2009-2010	-0.450 (p=0.02)	-0.303 (p=0.14)
2010-2011	0.458 (p=0.46)	0.400 (p=0.05)
2008-2009	-0.079 (p=0.71)	-0.129 (p=0.54)
2011-2012	-0.071 (p=0.74)	0.708 (p<0.001)

The 2014-2015 flu season was most highly correlated with the 2012-2013 flu season, showing a strong positive linear relationship for both physician visits (r=0.655) and emergency department visits (r=0.82). The 2009-2010 season had a strong negative correlation for physician visits (r=-0.45) and moderate negative

correlation for emergency department visits (r=-0.303). The 2008-2009 flu season showed no association with the 2014-2015 season. The correlation coefficient for physician visits and emergency visits generally followed the same trend, except for the 2011-2012 season. For the 2011-2012 flu season, physician visits were not correlated with physician visit data from 2014-2015 (r=-0.071). But, the emergency department visit data for 2011-2012 showed a very strong positive correlation (r=0.708) with emergency department visits for 2014-2015.

Aim 2

Aim 2 assessed the usefulness of Google Flu Trends in detecting ILI activity in Maryland. The level of linear association between DHMH data, represented by physician visits and emergency department visits and Google Flu Trend data varied. Results are presented in Table 3 in ranked order according to physician visits. Unlike the results from aim 1, the correlation between Google and DHMH surveillance data always had a positive relationship and the lowest level of correlation still represented a moderate relationship between the data sources.

Table 3: Linear relationship between Google Flu Trends and DHMH data for flu seasons 2008-2015

	Physician Visits	Emergency Department Visits
2009-2010	0.952 (p<0.001)	0.980 (p<0.001)
2010-2011	0.902 (p<0.001)	0.965 (p<0.001)
2014-2015	0.897 (p<0.001)	0.947 (p<0.001)
2013-2014	0.874 (p<0.001)	0.967 (p<0.001)
2012-2013	0.862 (p<0.001)	0.974 (p<0.001)
2008-2009	0.745 (p<0.001)	0.393 (p=0.02)
2011-2012	0.394 (p=0.02)	0.724 (p<0.001)

The weakest correlation was seen in the 2011-2012 flu season for physician visits (r=0.394), and the 2008-2009 flu season for emergency department visits (r=0.393). Apart from the 2008-2009 flu season, the relationship was consistently stronger between Google Flu Trends and emergency department visits. The strongest correlation was observed for the 2009-2010 flu season for both physician (r=0.952) and emergency department visits (r=0.980).

Aim 3

Aim 3 examined if tweets from a symptom based keyword combination were correlated with Google Flu Trends and DHMH influenza surveillance data to see if tweets could be a used as mechanism for influenza surveillance. The fully cleaned Twitter dataset had a very strong correlation (r=0.98) with the Twitter dataset that was cleaned for re-tweets and multiple tweets from the same user (referred to as the raw dataset). However, when calculating the Pearson's correlation coefficient between Twitter data and other sources of influenza surveillance, the fully cleaned dataset had a stronger relationship with all other sources (see Table 5). The raw dataset contained 18,112 tweets. Only 0.97 of the tweets contained geo-coded information. Due to the lack of geo-coded tweets and tweets containing Maryland location identifiers, no separate analysis was done comparing Maryland tweets to the full dataset. After cleaning the data to include only tweets suspected to have occurred in the United States, the sample size was reduced to n=12,268. 67.7% of the tweets returned for keywords "fever AND (cough OR sore throat)" were determined to have occurred in the United States based upon the location determination system developed in this study. From the cleaned dataset, only 952, or 7.8% of tweets

contained the words influenza or flu within the full tweet text. An example of some tweets in the final dataset can be found in Table 4, some of the examples show that while most tweets focused on experiencing symptoms, some noise still existed in the dataset.

Table 4: Examples of tweets for keywords fever AND (cough OR sore throat)

Please pray for healing. I have a bad fever and super sore throat.
why would u come to school w a fever, stuffy nose, sore throat, and aching body?
High fever and sore throat and all I want is a chocolate frosty
This sore throat, fever, runny nose, and back pains are already calling for a great night at work! ~feeling miserable~
#WheatgrassJuice can be used for treatment of respiratory tract complaints, including the common cold, cough, fever, and sore throat.
Fever, chills, sore throat...where did this come from? Is February over yet? #IHateFebruary #WorstMonthOfTheYear
Going to school with a fever and sore throat sucks):
Way to start my birthday month! sore throat, chills, headache, I feel the fever coming!!!! Google scares me
What is swine flu?C)Symptoms similar to those produced by standard, seasonal flu - fever, cough, sore throat, body aches and chills

Table 5: Pearson's correlation coefficients for cleaned and raw Twitter data and Google Flu Trends with DHMH surveillance data for the 2014-2015 flu season

	Cleaned Twitter Data	**Raw Twitter Data**	**Google Flu Trends**
Physician Visits	0.675 (p=0.001)	0.593 (p=0.006)	0.897 (p<0.0001)
Emergency Department Visits	0.642 (p=0.002)	0.530 (p=0.02)	0.947 (p<0.0001)
Lab Confirmed Cases	0.616 (p=0.004)	0.494 (p=0.03)	0.927 (p<0.0001)
Google Flu Trends	0.642 (p=0.002)	0.536 (p=0.01)	1.00

Results show that tweets had a strong positive relationship with all other sources of surveillance data. Pearson's correlation coefficients for frequency of ILI activity per week ranged from r=0.616 with laboratory confirmed cases to r=0.675 with physician visits (Table 5). Tweets had a lower correlation with all sources of DHMH influenza surveillance data than Google Flu Trends for the 2014-2015 flu season. It is interesting to note that Twitter and physician visit data lacked a strong peak in activity, as is usually seen during the flu season and as can be observed in the other forms of surveillance data, see Figure 3.

Sub-aim 3.1:

No racial indicators were included in the Twitter dataset and therefore no separate analysis could be performed to investigate whether or not sub-portions of the population are being under or over represented in the sample.

Sub-aim 3.2:

Tweets had a strong positive association with laboratory confirmed Influenza cases. However, the Pearson's correlation coefficient between tweets and laboratory confirmed cases was the lowest compared to the other surveillance sources analyzed.

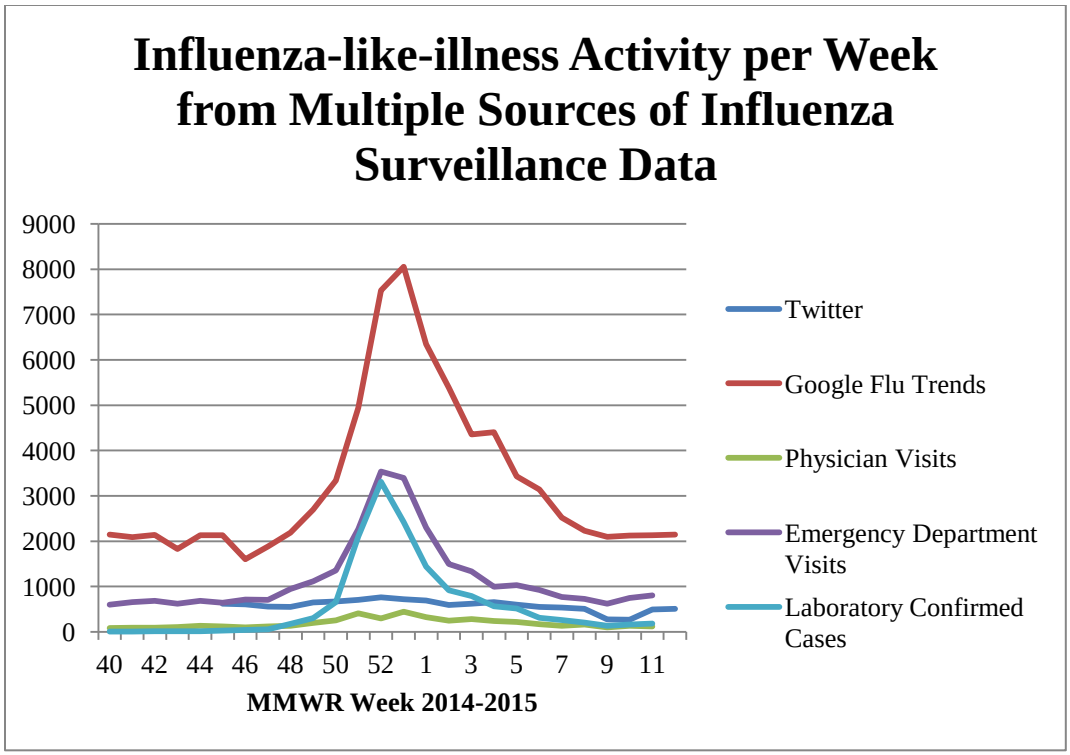

Figure 3: Graphical representation of influenza-like-illness activity from all surveillance sources used in this study

Chapter 4: Discussion

Aim 1

Aim 1 was performed to assess the linear relationship between the 2014-2015 flu season to past flu seasons. Results show that the 2014-2015 flu season was only comparable to three other seasons (two moderately, one strongly). Therefore, the results assessing the relationship between Twitter and DHMH influenza surveillance data is not generalizable to all flu seasons and may be more or less correlated with each season based on specific characteristics of that flu season. Differences in relationship between the 2014-2015 flu season with other flu seasons could be due to severity of the most prominent strain and when activity becomes more widespread. For instance, the 2014-2015 season was expected to be more severe due to a vaccine mismatch with the circulating influenza strains (25). This could be a reason why the 2014-2015 flu season had a low correlation with past seasons. The 2014-2015 flu season had the weakest relationship with the 2009-2010 and 2008-2009 flu seasons, which may be due to unique flu activity resulting from the 2009 Swine Flu (H1N1) pandemic (26). The 2012-2013 flu season was the most highly correlated with the 2014-2015 flu season. According to DHMH's flu season summary, the 2012-2013 flu season was the most active season since the 2009 H1N1 pandemic (27). Similarities of the 2014-2015 and 2012-2013 seasons are AH3 as the most prominent strain, and being an active flu season (23,25,27).

Aim 2

Google Flu Trend data were compared to DHMH physician and emergency department visits from flu seasons 2008-2015 in order to determine if Google has been useful in tracking influenza activity in Maryland. The magnitude of Google data compared to the other surveillance sources demonstrated that far more people seek information than care, and confirms that using Google Flu Trends for influenza surveillance provides information on cases that would normally be missed in surveillance relying only on people accessing healthcare (3,7,11). Results from Aim 2 were consistent with previous studies which showed an initial low correlation with clinical data at the beginning of the 2009 Swine flu pandemic, but that changes to the algorithm used in Google Flu Trends drastically improved the correlation between official data and Google Flu Trends for the remainder of the pandemic (7). Pearson's correlation coefficients between Google Flu Trends and DHMH data for the 2008-2009 season was ranked second lowest for physician visits, and lowest for emergency department visits while the 2009-2010 season had the highest correlation coefficient for both physician and emergency department visits.

The results of the linear relationship between Google and DHMH surveillance data were different than expected. Since Google revises the algorithm used to track flu activity it was expected that the most recent years would have the highest correlation coefficients. Variation may be due to differing characteristics of each flu season or the current algorithm may be perfected to pandemic H1N1 conditions. Google states that flu trend data should be interpreted as ILI cases per 100,000 physician visits (28). Interestingly, in this study apart from the 2008-2009 season,

Google data were consistently more highly correlated with emergency department visits. Based on the data in this study, in Maryland, the same population that uses Google to search their symptoms and influenza information might also be more likely to visit the emergency department rather than a physician's office for care. However, it is hard to differentiate who is represented and how different groups use Twitter.

Aim 3

In this study tweets were found to be positively associated with influenza surveillance data on physician visits, emergency department visits, and laboratory confirmed cases, as well as with Google Flu Trends. This study went beyond methods used in previous studies researching social media for influenza surveillance and took a different approach to better capture incident data. The higher Pearson's correlation coefficients reported in previous studies using keywords such as flu and influenza were likely heavily influenced by noise produced by tweets from public health organizations, news, and tweets related to flu vaccines. Since only 7.8% of tweets returned on ILI symptoms contained the words flu or influenza this provides further evidence that the use of flu and influenza as keywords for disease surveillance fails to identify the majority of self-reported ILI cases.

A sub-aim of this study was to explore characteristics to determine if certain sub-portions of the population were being under or over represented. Twitter is more popular with certain portions of the population, such as college students. However, it is becoming more diverse; a larger percentage of online African Americans use Twitter than online Hispanics or Whites (15). It is possible that the sample of tweets

used could be representative of all Twitter users and therefore a relatively heterogeneous sample. But, since there are no racial indicators on Twitter profiles, no comment can be made with certainty on the racial identities of those persons generating the tweets used in this study.

There were many challenges in using Twitter data for research purposes. No standardized method for determining location of users or tweets has been developed. Even with a flow chart guiding decisions on a user's location, data cleaning was a time consuming endeavor. This limits the ability to use Twitter data in public health settings due to time constraints. However, this may be overcome by using tweets on ILI that occurred world-wide, represented in this study by the raw dataset. The raw and cleaned dataset had a very strong positive correlation. While the relationship between raw tweets and DHMH surveillance data was weaker, there was still a positive association. So, the raw dataset can provide a rough estimation of activity, but ultimately the cleaned dataset is the best choice when using tweets for disease surveillance. Research is currently being conducted on algorithms that estimate the location of Twitter users and tweets (16). While time constraints currently exist, this limitation may be overcome in the near future with continued research and development.

While results of this study show Twitter is correlated with DHMH data there was no evidence that Twitter or Google Flu Trends showed increases in flu activity earlier than other surveillance sources. The main advantage of Google Flu Trends and Twitter for influenza surveillance is being able to access real-time data. Activity reports produced by DHMH were released a full week after the week being reported.

Even after this delay in reporting, often there was still missing data, resulting in a two week delay in obtaining complete ILI surveillance data. Public health officials themselves may not have to wait the entire two weeks to view surveillance activity. But, they are still limited by how many and how quickly physicians' offices and hospitals report ILI visits. So, while Google Flu Trends and Twitter might not show activity increasing earlier than DHMH surveillance, the data can be accessed sooner which is important for emergency management and public health officials preparing for and responding to outbreaks.

The interconnectedness of our world means that influenza outbreaks occurring across the country, or even world, can easily spread to Maryland. Not only can the data from Google Flu Trends and Twitter be accessed sooner, but an additional benefit of using these surveillance methods is being able to track activity outside of a health department's jurisdiction. Increases in influenza activity occurring in other parts of the country can help preparedness efforts for local health departments.

Limitations

There were a variety of limitations in this study. The method developed to determine the location of tweets has not been validated, and there is currently no standardized method that exists. This resulted in a time consuming data cleaning process that limits the application of Twitter outside of research settings, unless automated tools are developed to streamline this process. Since there were no racial/ethnic identifiers, the representativeness of the tweets cannot be verified. It is possible that the dataset could be over or under representing certain sub-groups, and therefore not representative of the entire population. Lastly, Google Flu Trends and

DHMH data were collected from Maryland, while tweets were collected from the entire United States. This means there is a difference in the base populations used in this study. However, it is hypothesized that the correlation would increase if only tweets from Maryland were used. Subsequently, the true correlations between tweets, DHMH, and Google Flu Trends might be higher than the correlations reported in this study.

Conclusions

In general, Google Flu Trends and ILI symptom based tweets were positively correlated with current surveillance methods used by Maryland's Department of Health and Mental Hygiene. Since every flu season was found to be unique, the overall relationship between Google Flu Trends and tweets may vary year to year. In conclusion, the results of this study reinforce that influenza surveillance data should be gathered from a variety of sources in order to provide the greatest understanding of influenza outbreaks. These different sources of surveillance represent different portions of the population, such as those not seeking healthcare, and provide earlier access to data on influenza activity in order to best prepare for and manage an outbreak (7). Future work should focus on development of a tool which automatically collects tweets based on ILI keywords and cleans the dataset, application of internet and social media surveillance to other diseases, and standardized methods for determining location from Twitter data.

Definition of Terms

Bot: an application that is programmed to produce tweets

Geo-coded: contains a geographic reference point

Influenza-like-illness: illness with symptoms of fever, and cough and/or sore throat used to estimate influenza activity

Morbidity and Mortality Weekly Report (MMWR): Weekly series containing timely public health information prepared by the Centers for Disease Control and Prevention

Real time surveillance: surveillance that occurs at or very close to the onset of the disease

Re-tweet: a re-post of a tweet

Tweet: A message/post on Twitter, also referred to as Twitter messages

Twitter: Social media platform where users share 140 character messages called tweets

Bibliography

1. Barry J. The Great Influenza: The story of the Deadliest Pandemic in History. New /York, New York: Penguin Books; 2009. 546 p.

2. U.S. Government. National Strategy for Pandemic Influenza. 2005 Nov.

3. Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009 Feb 19;457(7232):1012–4.

4. Denecke K, Krieck M, Otrusina L, Smrz P, Dolog P, Nejdl W, et al. How to exploit twitter for public health monitoring? Methods Inf Med. 2013;52(4):326–39.

5. Chunara R, Andrews JR, Brownstein JS. Social and news media enable estimation of epidemiological patterns early in the 2010 Haitian cholera outbreak. Am J Trop Med Hyg. 2012 Jan;86(1):39–45.

6. Salathé M, Freifeld CC, Mekaru SR, Tomasulo AF, Brownstein JS. Influenza A (H7N9) and the importance of digital epidemiology. N Engl J Med. 2013 Aug 1;369(5):401–4.

7. Milinovich GJ, Williams GM, Clements ACA, Hu W. Internet-based surveillance systems for monitoring emerging infectious diseases. Lancet Infect Dis. 2014 Feb;14(2):160–8.

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

9. Brownstein JS, Freifeld CC, Madoff LC. Influenza A (H1N1) virus, 2009--online monitoring. N Engl J Med. 2009 May 21;360(21):2156.

10. Schmidt CW. Trending now: using social media to predict and track disease outbreaks. Environ Health Perspect. 2012 Jan;120(1):A30–33.

11. Harris JK, Mansour R, Choucair B, Olson J, Nissen C, Bhatt J, et al. Health department use of social media to identify foodborne illness - Chicago, Illinois, 2013-2014. MMWR Morb Mortal Wkly Rep. 2014 Aug 15;63(32):681–5.

12. Centers for Disease Control and Prevention. How Flu Spreads [Internet]. Centers for Disease Control and Prevention. 2013 [cited 2014 Nov 10]. Available from: http://www.cdc.gov/flu/about/disease/spread.htm

13. Corley CD, Cook DJ, Mikler AR, Singh KP. Using Web and social media for influenza surveillance. Adv Exp Med Biol. 2010;680:559–64.

14. Achrekar H, Gandhe A, Lazarus R, Yu S-H, Liu B. Predicting Flu Trends using Twitter data. 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). 2011. p. 702–7.

15. Duggan M, Ellison NB, Lampe C, Am, Lenhart a, Madden M. Social Media Update 2014 [Internet]. Pew Research Center's Internet & American Life Project. [cited 2015 Jan 20]. Available from: http://www.pewinternet.org/2015/01/09/social-media-update-2014/

16. Mahmud J, Nichols J, Drews C. Home Location Identification of Twitter Users. ACM Trans Intell Syst Technol ACM Trans Intell Syst Technol. 2014;5(3):1–21.

17. Graham M, Hale SA, Gaffney D. Where in the World Are You? Geolocation and Language Identification in Twitter. Prof Geogr. 2014;66(4):568–78.

18. 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.

19. Centers for Disease Control and Prevention (CDC). Overview of Influenza Surveillance in the United States [Internet]. 2015. Available from: http://www.cdc.gov/flu/weekly/overview.htm

20. Kontos E, Emmons K, Puleo E, Viswanath K. Communication Inequalities and Public Health Implications of Adult Social Networking Site Use in the United States. J Health Commun. 2010;15(Supplement):216–35.

21. Office of Infectious Disease Epidemiology and Outbreak Response Infectious Disease Bureau Prevention and Health Promotion Administration Maryland Department of Health and Mental Hygiene. Maryland Weekly Influenza Surveillance Actvitity Report [Internet]. 2014 Oct. Available from: http://phpa.dhmh.maryland.gov/influenza/fluwatch/SiteAssets/SitePages/Home/Weekly%20Influenza%20Report%202014-10-4.pdf

22. Morstatter F, Pfeffer J, Liu H, Carley KM. Is the Sample Good Enough? Comparing Data from Twitter's Streaming AP with Twitter's Firehose. ICWSM. 2013 Jul;

23. Maryland Department of Health and Mental Hygiene. fluwatch [Internet]. Department of Health and Mental Hygiene. 2015. Available from: http://phpa.dhmh.maryland.gov/influenza/fluwatch/SitePages/Home.aspx

24. Freelon D. Twitter geolocation and its limitations [Internet]. 2013 [cited 2014 Nov 11]. Available from: http://dfreelon.org/2013/05/12/twitter-geolocation-and-its-limitations/

25. Robert Roos. CDC's flu warning raises questions about vaccine match [Internet]. CIDRAP. [cited 2015 Apr 7]. Available from: http://www.cidrap.umn.edu/news-perspective/2014/12/cdcs-flu-warning-raises-questions-about-vaccine-match

26. CDC Novel H1N1 Flu | The 2009 H1N1 Pandemic: Summary Highlights, April 2009-April 2010 [Internet]. [cited 2015 Apr 7]. Available from: http://www.cdc.gov/h1n1flu/cdcresponse.htm

27. Maryland Department of Health and Mental Hygiene. Influenza in Maryland 2012-2013 Season Report [Internet]. Available from: http://phpa.dhmh.maryland.gov/influenza/fluwatch/Shared%20Documents/FINAL%20FLU%20REPORT%202012_13_9SEP13_Final.pdf

28. Google Inc. Frequently asked questions [Internet]. Google.org flu trends. 2011. Available from: http://www.google.org/flutrends/about/faq.html