OJPHI

ISDS 2018 Conference Abstracts

ISDS INTERNATIONAL SOCIETY FOR DISEASE SURVEILLANCE

Correlation of Tweets Mentioning Influenza Illness and Traditional Surveillance Data

Zachary Heth*[2, 1], Kelley Bemis[1] and Demian Christiansen[1]

[1]Communicable Diseases, Cook County Department of Public Health, Forest Park, IL, USA; [2]CSTE Applied Epidemiology Fellowship, Atlanta, GA, USA

Objective

To determine if social media data can be used as a surveillance tool for influenza at the local level.

Introduction

The use of social media as a syndromic sentinel for diseases is an emerging field of growing relevance as the public begins to share more online, particularly in the area of influenza. Several applications have been developed to predict or monitor influenza activity using publicly posted or self-reported online data; however, few have prioritized accuracy at the local level. In 2016, the Cook County Department of Public Health (CCDPH) collected localized Twitter information to evaluate its utility as a potential influenza sentinel data source. Tweets from MMWR week 40 through MMWR week 20 indicating influenza-like illness (ILI) in our jurisdiction were collected and analyzed for correlation with traditional surveillance indicators. Social media has the potential to join other sentinels, such as emergency room and outpatient provider data, to create a more accurate and complete picture of influenza in Cook County.

Methods

We developed a JAVA program which included a customized geofence around suburban Cook County to collect tweets from Twitter's STREAM application programming interface (API) (available at https://github.com/FoodSafeCookCo/TwitterStream-Program). The JAVA program looked for tweets within the geofence or for tweets with a profile location naming a suburban Cook County municipality and copied them to a text file if the tweet contained "flu" or "influenza". Captured data included the user's Twitter handle, Tweet text, Tweet time and date, x and y coordinates (if available), and profile location. Tweets were then manually reviewed to determine if they met the following criteria: 1) language indicated the user was recently ill with influenza; 2) user appeared to reside in CCDPH jurisdiction. Tweets meeting these criteria were included in the analysis. Tweets were aggregated by MMWR week and analyzed for correlation, using Pearson methods (data were normal), with two traditional surveillance sources: 1) the percent of visits to all suburban Cook County emergency departments for ILI as reported to the Cook County Electronic Surveillance System for the Early Notification of Community-based Epidemics (ESSENCE), and 2) the percent of laboratory specimens testing positive for influenza at seven local sentinel laboratories. Analysis was performed in Excel 2013 and SAS 9.4.

Results

From MMWR week 40-20, 113 tweets indicating influenza-like illness were collected within Cook County's jurisdiction. Due to technical issues with the program, data were not collected for weeks 52, 2, and 17-19. Correlations were compared for the percent of laboratory specimens testing positive for influenza (LSL) and percent of visits to emergency departments for ILI (EDILI) to the total number of tweets per MMWR week. A strong correlation exists between LSL and EDILI r=0.92 (p-value<0.0001) indicating the traditional sentinels have a strong positive relationship. The correlation between number of tweets and LSL was 0.46 (p-value =0.0138), indicating a moderate positive relationship. Correlation between number of tweets and EDILI was similarly moderate, r=0.52 (p-value=0.0049). Correlations to EDILI stratified by age (0-4, 5-17, 18-64, 65+) also showed a moderate positive relationship (range 0.50 to 0.55, all p-values < 0.01). Twitter use peaked one week before the recorded peak of other surveillance indicators. When Twitter counts were shifted one week to align the peak in tweets with the peak of the influenza season, the correlations were 0.54 for LSL and 0.61 for EDILI (p-value=0.0034 and 0.0007, respectively).

Conclusions

Overall, the tweets collected had a moderately positive relationship with the severity of influenza activity in Cook County. When the data were transitioned to match peaks, there was an increase in the correlations' strength for both LSL and EDILI. This data indicates that publicly shared social media data may be an underutilized source of syndromic data at the local level, potentially capable of predicting seasonal influenza peaks before traditional data sources. Other jurisdictions may consider using the open source program created by CCDPH to determine the relationship of influenza related social media to their own local influenza surveillance data. For the 2017-2018 influenza season, we established a redundant system for tweet collection in case one of the systems goes down. Exploring machine learning (in place of manual review) to detect tweets indicating illness is also a promising avenue to simplify data collection and cleaning. Data will be collected using the same system for the 2017-2018 influenza season and correlations re-evaluated with more complete data.

Keywords

Twitter; influenza; syndromic surveillance; ESSENCE; social media

Acknowledgments

This study/report was supported in part by an appointment to the Applied Epidemiology Fellowship Program administered by the Council of State and Territorial Epidemiologists (CSTE) and funded by the Centers for Disease Control and Prevention (CDC) Cooperative Agreement Number 1U38OT000143-04.

***Zachary Heth**
E-mail: zheth@cookcountyhhs.org

ISDS Annual Conference Proceedings 2018. This is an Open Access article distributed under the terms of the Creative Commons Attribution-Noncommercial 3.0 Unported License (http://creativecommons.org/licenses/by-nc/3.0/), permitting all non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.