Information Systems Frontiers (2019) 21:743–760
https://doi.org/10.1007/s10796-018-9893-0

Social Media for Nowcasting Flu Activity: Spatio-Temporal Big Data Analysis

Amir Hassan Zadeh[1] · Hamed M. Zolbanin[2] · Ramesh Sharda[3] · Dursun Delen[3]

Published online: 5 January 2019

Abstract
Contagious diseases pose significant challenges to public healthcare systems all over the world. The rise in emerging contagious and infectious diseases has led to calls for the use of new techniques and technologies capable of detecting, tracking, mapping and managing behavioral patterns in such diseases. In this study, we used Big Data technologies to analyze two sets of flu (influenza) activity data: Twitter data were used to extract behavioral patterns from a location-based social network and to monitor flu outbreaks (and their locations) in the US, and Cerner HealthFacts data warehouse was used to track real-world clinical encounters. We expected that the integration (mashing) of social media and real-world clinical encounters could be a valuable enhancement to the existing surveillance systems. Our results verified that flu-related traffic on social media is closely related with actual flu outbreaks. However, rather than using simple Pearson correlation, which assumes a zero lag between the online and real-world activities, we used a multi-method data analytics approach to obtain the spatio-temporal cross-correlation between the two flu trends and to explain behavioral patterns during the flu season. We found that clinical flu encounters lag behind online posts. Also, we identified several public locations from which a majority of posts initiated. These findings can help health authorities develop more effective interventions (behavioral and/or otherwise) during the outbreaks to reduce the spread and impact, and to inform individuals about the locations they should avoid during those periods.

Keywords Business analytics · Big data · Public health · Social media · Behavioral analytics · Location analytics

1 Introduction and Motivation

Surveillance systems have long been the cornerstone of public health efforts that are designed to address a wide range of public health needs such as detection (early identification) of epidemics (or outbreaks) and control of infectious diseases. Such systems can inform policy decisions by identifying risk factors for disease and targets for preventive healthcare (Richards et al. 2014). Traditional surveillance systems have mainly relied on data reported by medical institutions, which oftentimes involve a very long data processing, thus increasing the uncertainty of decision making and more importantly delaying the effective intervention (Fang and Chen 2016). Furthermore, the failure of public health agencies to detect and manage emerging infectious disease is generally attributed to limitations of traditional surveillance systems for outbreak management (Y.-D. Chen et al. 2011). However, with the recent explosive growth in data availability as well as the ability to record collective behavior through digital data, social media, connected sensors and embedded systems (the so–called Internet of things), more efficient and reliable surveillance systems can be built in order to monitor the health status of a population at a given time. For instance, disease outbreaks can be detected with the help of a surveillance system based on query logs of search engines to determine in a timely manner where, when, and how health resources are allocated in order to achieve the best outcomes (Fang and Chen 2016).

✉ Amir Hassan Zadeh
Amir.zadeh@wright.edu

Hamed M. Zolbanin
Hmzolbanin@bsu.edu

Ramesh Sharda
Ramesh.sharda@okstate.edu

Dursun Delen
Dursun.delen@okstate.edu

[1] Department of Information Systems and Supply Chain Management, Raj Soin College of Business, Wright State University, Dayton, OH, USA

[2] Department of Information Systems and Operations Management, Miller College of Business, Ball State University, Muncie, IN, USA

[3] Department of Management Science and Information Systems, Spears School of Business, Oklahoma State University, Stillwater, OK, USA

Emerging infectious diseases are of paramount concern to public health officials and governments throughout the world. The susceptibility of people to infectious diseases has been evidenced by the emergence of HIV/AIDS in the late seventies, pandemic Swine flu in 2009, the H3N2 epidemic during the 2012-2013 winter season, and more recently, the Ebola virus disease worldwide, as well as the reemergence of Swine flu in India. In most countries, influenza outbreaks happen in different forms every year and invoke different consequences of varying impacts. The annual impact in the US has been estimated at an average of 610,660 undiscounted life-years lost, 3.1 million hospitalized days, 31.4 million outpatient visits, and a total of $87.1 billion in economic burden (Molinari et al. 2007). In order to minimize this cost, early detection and rapid development of behavioral interventions aimed at reducing the spread and impact of disease outbreaks are critical for public health.

The use of disease detection and surveillance systems are crucial in providing the necessary epidemiologic intelligence that health officials and clinical administrators rely on to properly enact preventative measures and assist in staffing and stocking decisions (Santillana et al. 2016). While several surveillance systems have been used for monitoring influenza virus activity, an effective disease surveillance is expensive and needs a formal public health network (Wilson and Brownstein 2009). Hospital admissions data (Congdon 2005), pneumonia and influenza mortality rates (Sebastiani et al. 2006), over-the-counter drug sales (Magruder 2003), syndromic/sentinel surveillance (Griffin et al. 2009), prescription pharmaceutical sales (Patwardhan and Bilkovski 2012), laboratory test isolation (Nunes et al. 2013), and emergency room visit rates (Corberán-Vallet and Lawson 2014) are among instances of traditional surveillance systems. However, none of these measurements can be deemed as the best single surveillance information source (Amorós et al. 2015), and such passive surveillance systems typically rely on data sources submitted to the public health authorities by networks of physicians, hospitals, laboratories, pharmacies, and other healthcare providers (if well run and efficient) and may not reflect behavioral patterns accurately.

The Centers for Disease Control and Prevention (CDC) is the US's leading public health agency that monitors influenza-like illness (ILI) activities by collecting data from medical institutions, collating reports and publishing them on a weekly basis. However, the lack of operational knowledge of reporting systems and real-time monitoring have resulted in a miscount of influenza cases and substantial lags (roughly 2 weeks) between the point at which an outbreak trend starts to happen and the point at which that data point becomes available in aggregate ILI reports (Amorós et al. 2015). These substantial lag times do not allow for optimal decision making, as public health decisions need to be based on more recent information. These issues can adversely affect the effectiveness of behavioral interventions targeted at reducing the spread and impact of outbreaks. Because of this, there is a glaring need for systems that can provide real-time influenza activity estimates. To have a proactive surveillance system, public health officials need to be forewarned at the earliest possible time to ensure effective behavioral interventions and shorten the course of outbreak complications. This has led to calls for more augmented surveillance systems capable of incorporating behavioral patterns, estimating the prevalence, and tracking the spread of influenza. The emergence and global spread of unknown variants of influenza in recent years prompted CDC and other public health agencies to call for rapid development of behavioral interventions aimed at promoting health-related behaviors and minimizing the spread and impact of pandemic outbreaks. Such recommended behaviors can be, for instance, keeping distance from flu-infected people or those who show ILI symptoms, such as sneezing and coughing, or reducing the time spent in crowded places during the peak flu season (Prati et al. 2011). More recently, CDC launched the "Predict the Influenza Season Challenge Competition" to encourage researchers to characterize the flu season that occurs each year using digital data from various sources such as social media, search queries, or other Internet data.[1] In particular, social media data can be crucial to identify more real-time behavioral patterns during the flu season and to promote health intervention efforts.

Today, in the era of online communication and social media, thousands of online services and mobile applications have yielded and will continue to yield a vast amount of user-generated and location-based content every day. Many people seek information about health online, and often, use social media to share their thoughts, feelings, and experiences publicly with their friends (Nguyen et al. 2015). The availability of such patient-generated health data on the Internet, with varying quality and legitimacy, has provided a new means for not only early detection of infectious diseases, but also surveillance of patient behaviors from the early recognition and interpretation of symptoms to the representation of disease and treatment-seeking. During a new disease outbreak, not only that affected communities can benefit from social media by looking for known symptoms, preventative measures and treatment related information, but also health organizations can analyze millions of messages exchanged on social media platforms about disease signs and symptoms, transmission mediums, death reports etc. for quick response and and implementation of intervention strategies (Rudra et al. 2018). Although an Internet-based surveillance system is only limited to people who seek health-related information on the Internet, we can sense the simultaneous presence of real and "digital" outbreaks from such seemingly social networking

[1] http://www.cdc.gov/flu/news/predict-flu-challenge.htm

data; we can read the city's public health status in real-time, determine infectious hotspots, and highlight "must-know facts" for people visiting these areas.

With such geo-referenced and time-stamped social media data at hand, patient-generated health data can be synthesized and mashed up with large sets of clinical and medical data to gain valuable insights for patients and healthcare providers, gain insights into complex infectious disease conditions, and anticipate public health outbreak crises that ensue. As a result of the presence of Big Data in healthcare, digital epidemiology has emerged as a major field of research that investigates how "big data" can be used to better understand, detect, monitor, and address public health problems (Young et al. 2014). This field has the potential to significantly aid public health officials as far as launching vaccination campaigns, allocating resources, and other strategic measures used to fight the spread of infectious diseases.

Over the last few years, several innovative web-based flu surveillance systems have been proposed to improve the overall efficiency of public health monitoring. Internet surveys (Vandendijck et al. 2013), search queries (Dukic et al. 2012; Fang and Chen 2016; Santillana et al. 2015), and microblogging (Signorini et al. 2011) have been used to discover public health seeking patterns and transform them into flu activity. Perhaps, the development of the Google Flu Trends (GFT) (Ginsberg et al. 2009) has been the most groundbreaking innovation of the web-based surveillance systems, which later led to an increasing focus on harnessing Internet-scale data for public health surveillance and planning. GFT was shut down in August of 2015, creating a need for a new, reliable method to replace it. GFT allowed for significant progress in the field of digital epidemiology, as both Google and other researchers have been able to recommend updates and improvements to GFT (Copeland et al. 2013). A number of these updated models have shown notable improvements in areas such as inputting historical flu activity and dynamically recalibrating models, as to ensure the usage of the most recent clinical information and adapt to changing behavior within the population (Lampos et al. 2015; Wagner et al. 2018).

Several studies have shown that social media can also be used as an enhanced method for faster detection of influenza outbreaks (Broniatowski et al. 2013; Signorini et al. 2011; Louis and Zorlu 2012). Most of these studies make the assumption that digital surveillances are substitutes for, rather than supplements to, the traditional surveillance data sources. However, these sources of information provide just indirect estimates ("nowcast") of the ILI cases in the population and do not provide an absolute alternative to existing surveillance systems. Rather, they can improve the forecast by informing other models of the state of the observed conditions in real-time; thus, allowing the surveillance system to operate at its best. Therefore, new developments should be mostly focused on exploring the ways to integrate web-based data sources into existing forecasting/surveillance systems, rather than retrospective analyses of performance (Aslam et al. 2014; Lazer et al. 2014; Milinovich et al. 2014). More recently, several studies advocated use of hybrid systems combining information from traditional surveillance and big data sources such as search queries, social media and health forums (Al-garadi et al. 2016; Davidson et al. 2015; Simonsen et al. 2016). Applying advanced GIS and machine learning techniques to social media data can provide opportunities to understand behavioral patterns and characterize the spread of influenza outbreaks, in real-time and at different geographical scales (Allen et al. 2016).

Prior works on flu trend analysis on Twitter have mostly focused on temporal information, and how well the Twitter data fits the clinical data. However, the spatio-temporal dynamics of the flu outbreak needs to be further studied. In this paper, we focus on flu outbreaks in U.S. and examine that whether social media data can be used to effectively identify real-time behavioral patterns during the flu season, and if so, can this lead to more timely promotion of health intervention efforts. Specifically, this study seeks to answer the following research questions: 1) Do flu-related social media posts precede or help identify the flu outbreak? 2) Do the spatio-temporal dimensions of social media data exhibit evidence of infectious hotspots or signal the presence of flu viruses in the vicinity of certain points of interest?

While prior studies have analyzed the flu trend at the aggregate level, our work uses a fine level of granularity (i.e., inpatient/outpatient encounter) and employs a series of correlations and time series-based models to examine the spatio-temporal relationship between the incidence of flu and social media. In particular, the point processes technique introduced in this study is capable of tracking behavior at lower levels of granularity to test the correlation of Twitter data with clinical data. Despite the increasing emphasis on using the massive amounts of data available through social networks to employ the potential of big data technology in behavioral research, few studies have used big data analytics as a tool to support decision making in public health decisions, especially for pre-emptive actions prior to disease outbreaks. Hence, in response to the call for more examples on the use of big data for analytics in health behavior research, in this paper, we created large datasets of flu-related activities from Twitter and Cerner medical records and built analytical models to understand patterns and trends in these datasets. The data management approach presented in this paper utilizes a Hadoop infrastructure to develop a social media analytics application for flu spread monitoring. We demonstrate an example of how big data technologies such as Hadoop, HDFS, MapReduce, Hive, Flume or BigR are developing into a more advanced and interactive form of application development.

We used a big data technology framework to build a model that employs Twitter data to track influenza activities in the

US. Our analytics initiative comprised temporal and spatial big data analyses. In the temporal analysis, we analyzed whether Twitter data could indeed be adapted for the nowcasting of influenza outbreaks. In spatial analysis, we mapped flu outbreaks to the geo-spatial property of Twitter data to identify influenza hotspots. In addition, we used lead-lag theory, also known as the sequence theory, to investigate the relationship direction between online posts and actual cases of the flu, as online discussions may result from, precede, or reflect flu outbreaks. The lead-lag effect theory, especially in economics, is based on the principle that changes in the business are not simultaneous, but successive. It describes the situation where one (leading) variable is cross-correlated with the values of another (lagging) variable at later times (Wang et al. 2013). Guided by this theory, we demonstrated the application of point processes as an integrated framework to investigate lead-lag relationship and spatio-temporal cross-correlation between two trends in a location-based social network context. We tested not only the contemporaneous relationship of flu activity on Twitter and clinical health records, but also whether one preceded the other. To the best of our knowledge, there is no prior research on using point process as its modeling framework for characterizing the spatial and temporal relationship between the disease outbreaks and online social networks. One obvious contribution of this study, therefore, is its roles as a proof-of-concept for the use of spatial big data analysis and mapping methods to supplement traditional disease surveillance systems. We utilized the analytical and visualization capabilities of GIS to enhance our understanding of early disease spreading patterns using location-enabled social media, otherwise hard to discern from patient medical records in near real-time. We further analyzed the geographic patterns to identify influenza public hotspots. In addition, the geostatistical measures were used for auto- and cross-correlations to examine the spatiotemporal distribution of flu activity. The study provides theoretical, empirical, and methodological advances that help policymakers use location-based social media platforms as a source of spatio-temporal information for spread of various diseases.

The remainder of this work is structured as follows. In the next section, we provide an overview of recent work in related research areas. Subsequently, we provide objectives and outline of our research according to the identified gap. In the following section, we briefly review the theoretical background of our research, followed by a section that covers the data and research methods employed in this study. The application of the statistical and data mining models to evaluate the explanatory power of location-enabled social media is presented in the next section. Therein, the flu mapping using geo-spatial social media activity is also illustrated. The penultimate section presents the findings of our research and provides an outlook on future research. The paper concludes a summary of the research method the findings.

2 Theoretical Background

In this section, we review theories that explain how human behavior properties impact the propagation of information on social networks. Among all the theories that exist in IS research, theories of critical mass, social exchange, and attachment motivation conform to our observations/findings and therefore, are brought up as potential determinants of human behaviors in information propagation phenomena online.

2.1 Critical Mass

The critical mass theory proposes that social movements start once a specific threshold of participants or actions involved in a collective behavior is crossed (Oliver et al. 1985). Previous studies have investigated the role of critical mass in adoption of microblogging platforms and its influence on the intention for the cascading behavior such as re-tweeting (Shi et al. 2014). It is asserted that the perceived critical mass impacts the collective or cascading behavioral intention in social networks.

A disease outbreak needs a critical mass of infected patients and a critical mass must be present for patient one to become infected and susceptible to the infection. This threshold can be inferred from social media activities. When a large number of people are sharing content related to flu on social media platforms, that is likely a critical mass for an early-warning surveillance system of a change in influenza activity in a city or state.

2.2 Social Exchange

Social exchange theory states that people engage in a process of social exchange expecting returns for their input (Homans 1958). In the IS literature, content sharing is considered a social exchange mechanism that is formed with the involvement of three parties: the creator, the sharer, and the targeted audience of the content. The literature has emphasized the rewarding aspects of social interactions for humans and that the perceived reputation enhancement is accounted as a key factor in explaining human behavior in information sharing. Social exchange theory provides additional insights into the factors why people talk about health conditions on social media platforms. Social media is used as a means of social exchange in which a society exhibits social strain in connection with a disease and throughout this interchange, relevant health information is offered by the healthcare professional.

2.3 Attachment Motivation

The attachment motivation theory is used to explain online user's behavior in social networks due to an individual's desire for social interaction and social communication (Oliver et al.

1985). The attachment motivation theory sheds light on people's motivation to spread health-related information on social media, yet it has to be evaluated as a credible or reliable source of information.

2.4 Lead-Lag

The lead-lag effect is a concept of common practice in econometrics that has been used to describe the relationships that exist between a pair of spatial time series data items. Examples of this effect include the temporal relationship between social media (blog buzz) and sales (Dewan and Ramaprasad 2014), the spatial relationship between the utilization of information technologies and socioeconomic factors (Pick et al. 2015), and the spatiotemporal relationship between criminal activity and socioeconomic variables (Toole et al. 2011). In time series, this effect is commonly referred to as "cross-correlation." The notion of lead-lag relationship can be linked to the concept of Granger causality and its implications to understand which variable is more likely to precede the other (Granger 1969). In theory, social media and electronic word-of-mouth (eWOM) activities may not only result from real-world dynamics (e.g. (Duan et al. 2008; Lymperopoulos and Ioannou 2015)), but also may inform it in due course of time (e.g. (Dewan and Ramprasad 2009)). Statistical measures comprising both spatial and temporal cross-correlation processes can provide a better estimate of such relationships (Y. Chen 2015; Ma et al. 2006). In summary, all the theories discussed above are complementary and useful for us to understand the nature of the relationship between the social media activities and flu behavior.

3 Data and Methods

The exponential growth of big data technology in today's world has changed the way we manage and process data. It has led to an architectural paradigm shift in data movement. Instead of bringing the data to computation, the new architecture pushes computation to the data. The move from 'scale-up' to 'scale-out' architecture requires a very different approach to data storage and processing. Big data technology as a viable solution consists of a set of highly scalable tools and techniques capable of moving beyond terabytes and even exabytes of data. Hadoop as an integral part of big data analytics has fundamentally changed the economics of storing and analyzing information. The high rate of unstructured data production on social media makes it appealing to use Hadoop in the production environment.

Digital data is massive with real information hidden within an immense amount of irrelevant data. Hadoop as an integral part of big data analytics has fundamentally changed the economics of storing and analyzing such huge amounts of semi-structured and unstructured data. Guided by design science theory (von Alan et al. 2004), the approach presented in this paper leverages a Hadoop platform to design and develop a social media-based surveillance system for tracking flu activity. We used IBM InfoSphere BigInsights as our big data analytics platform to derive underlying relationships in flu-related data sources, and to understand how well such information can help in achieving public health goals. BigInsights is the IBM's distribution of Hadoop. It combines an IBM-created open source query language called JAQL with the usual Apache Hadoop components such as Hive, HBase, Pig and Java MapReduce. The reason we used the Hadoop ecosystem is the power, the scalability and the functionalities it provides to our infrastructure to support what we do on the data. We ingested Twitter data using Apache Flume where many flows (queries) run in parallel (via Flume interceptors). The Hadoop's ability to efficiently collect large volumes of complex data in parallel, process and aggregate it with data from other sources via various applications available within the same package has made Hadoop a powerful system. All these assumptions necessitated the use of a Big Data Analytic platform (i.e. IBM BigInsights) in this paper to manage and process vast amounts of data that can, otherwise, be technically challenging to manage and analyze conveniently if traditional data management systems were used. As shown in Fig. 1, the proposed big data analytics methodology of this paper consists of four major modules: data processing, data manipulation, data modeling and forecasting.

3.1 Data Processing

Unlike prior studies that used the CDC data at the regional level on a weekly basis, the clinical data used in this study was acquired from Cerner Center Health Facts ™. The Cerner Data Warehouse (DW) is one of the largest de-identified, longitudinal electronic health records (EHR) databases with more than 58 million unique patients across the US. It contains comprehensive demographic, clinical, laboratory, pharmaceutical, admission and billing data that are geo-tagged, time-stamped and sequenced at a fine level of granularity (i.e., inpatient/outpatient encounter). More importantly, the Cerner DW includes both the patient's residential address and the facility address for every encounter. The final dataset is also identified at a fine level of granularity where the patient was admitted to a healthcare facility. At the time of data collection, the size of the Cerner DW was about 2.5 terabytes. We used big data technologies to extract 9.25 gigabytes from the DW, which roughly corresponded to over eight million flu-related medical visits. Sqoop was used to transfer this dataset to the Hadoop cluster.

The Twitter data crawler continuously streams tweets using the open-source Flutrack Twitter Streaming Application Program Interface (API) (Talvis et al. 2014) and forwards it

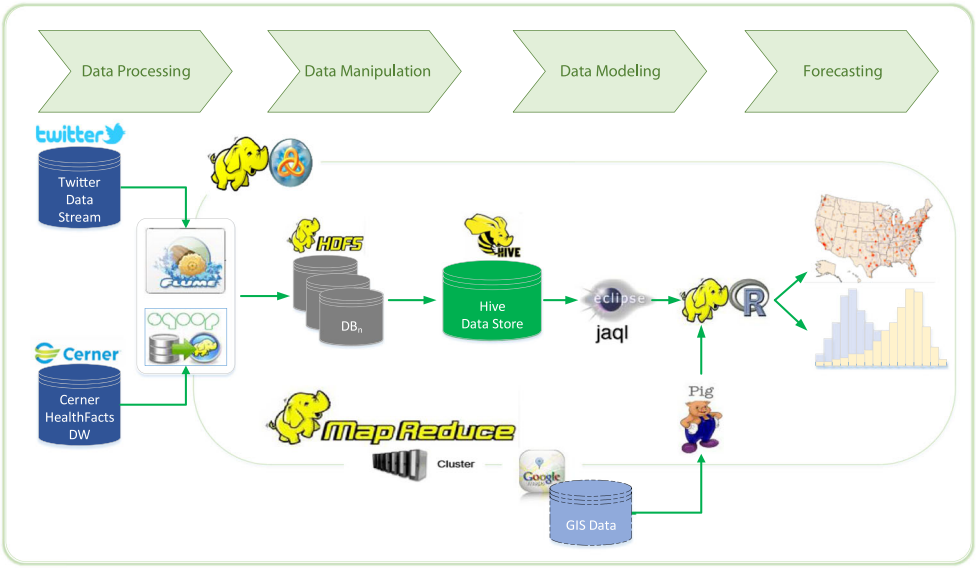

Fig. 1 High level architecture of the proposed system

to the Hadoop. A Flume program perpetually creates new files from the incoming tweet stream and stores them in the Hadoop Distributed File Systems (HDFS). The monitoring words used as tags were influenza and flu like symptoms. These included body or muscle aches, chills, cough, fever, headache, and sore throat. The Flutrack API filters out tweets with false or nonexistent location coordinates. Our Twitter dataset consists of over 26,000 public tweets that contained the influenza-like symptoms-related keywords generated by more than 20,700 unique users between May 06, 2013 and March 15, 2015. As a data storage system, we used HDFS, which is also a part of BigInsights Hadoop platform, to store the data collected from Twitter, as well as the Cerner Center Health Facts sub-dataset, where all the analyses were parallelized through MapReduce.

3.2 Data Manipulation

As is common with most data mining initiatives, data extraction, preparation, and cleaning stages constituted a significant amount of our work. Creating the flu dataset from the Cerner data repository included several steps. Cerner Health Fact data warehouse includes a database management system storing eight database Tables. A complete list of these tables along with their descriptions is provided in Table 1.

The encounter number is common among the tables. We joined all the fact tables and created one database to be used in our analyses. Since influenza-like symptoms are well known, we queried the Cerner health fact database to include those encounter records that were diagnosed with influenza, flu, fever, chills, headache, sneezing, sore throat, runny nose, cough, and dry cough. We removed all the duplicates and ensured that there is only one record per patient per flu activity in the table. Since there is one record per encounter number, a single person can have more than one record (visit) in this table over time. To make the dataset more manageable, we only included five columns: encounter ID, patient ID, admission date and time, location and diagnosis description. A list of the variables in the final Cerner Flu dataset, along with their descriptions and primary tables is provided in Table 2.

A large volume of flu-related tweets was collected from Twitter using Flutrack API and was ingested into a Hive table on the Hadoop cluster. Once the data was successfully imported, the JSON Query Language (JAQL) tool was used to manipulate and parse semi-structured data into a structured format. For each tweet, such information as username, timestamp, geographic location, and text were recorded. Twitter text is noisy, with linguistic errors and idiosyncratic style. We performed various cleaning and transformation processes such as stemming, spellchecking, and stop-word filtering to assure some reasonable quality. The Flutrack API itself classifies tweets as being related or not related to influenza (flu-positive/negative) (Chorianopoulos and Talvis 2016). Upon analyzing the collected tweets, we observed that some tweets were related to flu but did not report an infection. These tweets were not from influenza-infected users, but rather concerned with other flu-related activities such as fear of getting the flu, preventive vaccines (e.g. flu shots) and awareness of increased infections as opposed to infection. Similar to (Lamb et al. 2013), a further linguistic filtering process was designed to identify the tweets that were related to infections. We created a set of word class features (i.e. infection, possession, concern, vaccination, past vs. present tense, self vs. other) and trained a support vector machine (SVM) classifier to classify infection tweets. We used 1000 tweets as the training data, and another 200 tweets for testing. The SVM classifier demonstrated a classification accuracy of 94% as shown in Table 3 and was subsequently used to classify tweets that report infections. In

Table 1 List of tables included in the Cerner Health Fact Database

Table name	Description
Encounter facts	The "encounter fact table" contains all of the information that is specific to a certain visit. The encounter number is unique within this table.
Diagnosis facts	The "diagnosis fact table" has one record per diagnosis code, priority and type.
Procedure facts	The "procedure fact table" has one record per procedure code and priority.
Medication facts	Each row in the "medication fact table" has information about the pharmacy orders.
Laboratory facts	Each record in the "lab procedure fact table" has a different result.
Microbiology facts	Each row in "the microbiology fact table" has information about the microbiology orders and results.
Microbiology susceptibility facts	Each record in the "microbiology susceptibility fact table" has a different order, antimicrobial, and result/interpretation.
Clinical event facts	Each record in the "clinical event fact table" has a different event per event time and result.

addition, to validate the quality of the tweets (as being related to flu infection), 100 tweets were randomly sampled from the dataset in the previous step and were given to two medical experts for further evaluations. The experts were asked to label each of the tweets accordingly if they believed the tweets did or did not represent a flu infection. The experts approved 98% of the tweets to be related to flu infections. Table 3 summarizes the performance of the supervised classifier on the test dataset.

Next we analyzed tweet texts from the social network analysis perspective. We transformed the term-document matrix, built in the previous step, to a term-term adjacency matrix. The term-term adjacency matrix was used to build a network of flu symptoms. In summary, all of these efforts relied on natural language processing (NLP) algorithms to analyze tweet contents and classify them as being related or not related to flu. This process is supported by the attachment motivation theory. As the public perceive benefits from social media as a means of quick communication, individuals are more willing and motivated to share their health concerns or conditions with others on social media, especially during a crisis such as an outbreak of flu. However, their information should be evaluated for credibility. A list of the variables along with their descriptions is provided in Table 4.

In this study, the initial Cerner dataset contained the US flu incidences since 2009; however, our Twitter dataset included tweets from all over the world between 2013 and 2015. We created a subset of each dataset so that both subsets relate to the same period of time and location (i.e., US). This allowed us to compare the two subsets based on time and geolocation. Both datasets contained geo-tagged and time-stamped flu-related activities as specified in Tables 2 and 4. Consequently, the size of the datasets decreased to 1.5 GB for the Cerner data and to less than 1 GB for the Twitter data.

3.3 Data Modeling

In this work, all the statistical and data mining analyses of the data residing in HDFS were done through the R interface. The regular R installation is constrained by the size of the main memory and cannot scale up to large datasets. We used the package "BigR" released by IBM, to submit our R scripts over the data stored in HDFS. This package enabled us to benefit from the parallelism and scalability of HDFS, as well as in-database R analytics libraries to operate on big datasets that would otherwise not fit in memory (Lara Yejas et al. 2014).

We performed spatio-temporal big data analysis and visualization within a Geographic Information System (GIS). In the temporal analytics, we examined if Twitter data could indeed be adapted for nowcasting of influenza outbreaks. We tracked and compared clinical flu

Table 2 List of variables included in the Cerner Health Fact Flu Dataset

Variable name	Description
Encounter_ID	The visit identifier for the patient that this record is associated. This number is unique to the encounter (visit).
Patient_ID	The unique id used within the Cerner Health Facts Data Warehouse and created as a new person is introduced to the database. It is possible that the same "person" will have encounters.
Admitted_DT_TM	The date and time when the patient was admitted to the healthcare facility.
Location	The location (latitude and longitude) where the patient was admitted to the healthcare facility.
Diagnosis_Descript	The description associated with the ICD-9-CM diagnosis code.

Table 3 Classifier performance summary

Classifier	Precision	Recall	F-Measure
Support vector machine	0.938	0.905	0.921

encounters and flu-related activities on Twitter during the outbreaks. Rather than using the Pearson correlation, which assumes a zero lag between online and real-world activities, we used point processes analyses to obtain the temporal-spatial cross-correlations between the two trends. In spatial analysis, by the use of Google's maps API (an open-source API available to developers), we mapped flu outbreaks to geo-spatial property of Twitter data to identify influenza hotspots, i.e., public locations from which a majority of posts initiated. We assumed that flu-related social media traffic exhibits a congruence with actual flu outbreaks.

Spatial and temporal correlations fall into two categories: autocorrelation and cross-correlation. Cross-correlation is related to how the measures affect each other, using precedence and the contemporaneous relationship between them. Autocorrelation studies the relationship between one measure and itself, known as the intra-sample relationship. These types of correlation can be applied to both natural and human phenomena throughout a variety of fields, specifically in regard to GIS and their spatial analytical technology (Chen 2015). Figure 2 shows the roadmap from simple cross-correlation to the temporal-spatial cross-correlation introduced in this study.

Unlike prior studies that compare Twitter data with CDC data on a weekly basis, our Cerner Health Fact data is finely geo-tagged and time-stamped to the exact moment and location at which the patient record was created in the Cerner database due to an influenza-like illness. Thus, it allowed us to statistically examine the simultaneous relationship between actual flu incidence and social media activity, looking at both the temporal-spatial precedence and the contemporaneous relationship.

The dependence is inherent in many (if not most) spatial data and in most instances, spatial data can be viewed as a realization of a stochastic point process where the observed spatial pattern (or space-time pattern) yields information on the characteristics of that process (Anselin 2013). In this paper, we propose a flexible framework based on Hawkes point process capable of modeling such spatio-temporal patterns and specifying the spatial and temporal dependencies of flu outbreaks with a branching structure. We demonstrate that a spatio-temporal extension of Hawkes' point process, the ETAS (Epidemic Type Aftershock Sequence) model (Ogata 1988), is a suitable framework to model the cascading behavior of a contagious outbreak with conditions spreading infection to other locations. In addition, drawing on spatial-temporal dependence theories using point processes approach, we examine mutual and causal dependence between the two point processes under study, i.e., flu-related activities on Twitter and clinical flu encounters in Cerner medical records during the outbreak seasons.

A point process is a type of random process that can generate times and locations of events. Point process models have long been used to describe real-world phenomena occurring at random locations and/or random times. It is well-known that human activity is not uniformly distributed in space or time and the spatio-temporal dynamics of human interactions, which forms dynamic spatio-temporal hotspots, can be modeled using spatio-temporal point process models (Anselin 1989; Gonzalez et al. 2008; Mohler 2014). A point process is characterized by its conditional intensity function that allows us to describe the underlying dynamics of the process in a convenient way. The self-exciting point process was first introduced by Hawkes (1971) to structure temporal dependencies between earthquake events. Later, a spatio-temporal extension of Hawkes point process, the ETAS model, was formulated by Ogata (1988) to accommodate both temporal and spatial factors. The same framework was also used to study the link between behavior and spike patterns of neurons in order to understand brain functions (Brillinger et al. 1976). The following section provides more details about the point process method employed in this paper.

Model Development We built two point processes based on the Epidemic Type Aftershock Sequence (ETAS) model (one corresponding to flu activity on Twitter and the other to Cerner

Table 4 List of variables included in the Twitter Flu Dataset

Variable name	Description
Username	It is a name associated with the user who posted this tweet.
Time-stamp	It is a long integer that represents the number of seconds between the Unix Epoch (January 1 1970 00:00:00 GMT) and the time of Tweet generation.
Location	It indicates the tweet's location according to the spherical coordinates of longitude-latitude.
Text	It is a short text message limited to 140 characters in length posted by a user.

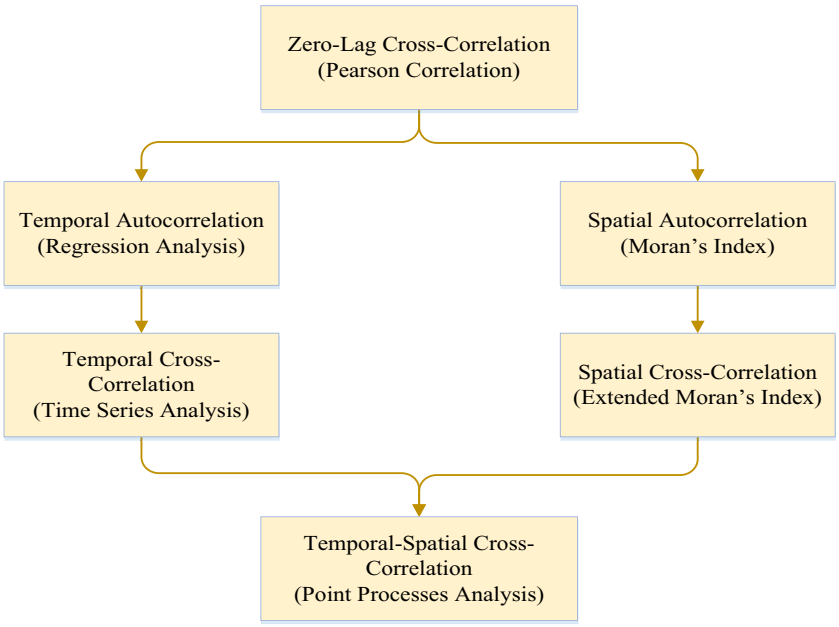

Fig. 2 The roadmap from simple cross-correlation to the spatio-temporal cross-correlation used in this study

flu data during the 2013-2014 flu outbreak). The conditional intensity of each of the ETAS spatio-temporal point process is formulated as a function of time and space coordinates:

$$\lambda(t,x,y|H_t) = \frac{E\left[N\left(dtdxdy|H_t\right]\right.}{dtdxdy} = \mu(t,x,y) + \sum_{i:t_i<t} g(t-t_i) f(x-x_i, y-y_i)$$

Where i is an index of flu activities occurred during the outbreak period, and $(x_i, y_i, \ t_i)$ denote the spatial location and time of event i, respectively. $H_t = \{(x_i, y_i, t_i : t_i < t\}$ represents the history of events prior to event i, and the expectation represents the number of events occurring in an infinitesimal space, $dt\ dx\ dy$. $\mu(x, y)$ denotes the background intensity corresponding to the baseline spatial and temporal patterns of outbreak. It corresponds to flu cases that were not necessarily caused by the flu outbreak, but exogenous factors (e.g. seasonality, vaccine availability, population density) independent of the history of the epidemic process. We assume the background events occur independently according to a Poisson process with their numbers varied by location and time.

The summation component of the ETAS model represents the epidemic proportion of the intensity process. It describes the proportion of events triggered by previous events. $g(t)$ is a power-law function that allows us to simulate the occurrence times of the triggered events and $f(x, y)$ is an exponential distribution of the locations of the triggered events. In other words, each of the (background) flu events elevates the risk of outbreak such that elevated risk spreads in space and time according to the kernel function $G(t, x, y) = g(t - t_i)f(x - x_i, y - y_i)$. The kernel function represents the intensity of triggered events (i.e. flu-infected people) from the background events. The power-law component of the kernel

$$g(t-t_i) = \frac{\beta}{(t+c)^p}$$

represents the temporal lag between the occurrence times of triggering and triggered flu events. The exponential component of the kernel

$$f(x-x_i, y-y_i; M_i) = \frac{e^{\alpha(M-M_0)}}{(x^2+y^2+d)^q}$$

describes the decay in intensity of triggered events in accordance with the spatial lag, which represents a distance between locations of a future event and a past event. M_i represents the magnitude of the event and implies how the event influences the occurrence of future events. In the context of seismology, M_i is associated with each event (t_i, x_i, y_i) and represents the magnitude of the earthquake. It is computed according to such scientific theories as the Gutenberg-Richer and the Omori laws. However, in our case, the functional form of how flu infection affect the virus spread is unknown. We, therefore, computed M empirically from the data by counting for each event the number of occurrences of consecutive events within the same region during a week. This is supported by the evidence that an adult infected with flu may be contagious up to 7 days after becoming sick. In addition, M_0, α and β are parameters that control the number of triggered events (i.e. flu-infected people), p and c are parameters that control the (power law) rate of decay according to the temporal lag and q and d are parameters that control the spatial behavior of the outbreak. These parameters are estimated via the maximum likelihood estimation (MLE) method. In the ML step, the vector of parameters $\theta = (\mu, \alpha, \beta, M_0, c, p, d, q)$

is computed by maximizing the log-likelihood function of the ETAS point process:

$$L(\theta) = \sum_{i=1}^{n} \log\lambda(t_i, x_i, y_i; \theta\ ) - \int_0^T \iint_S \lambda(t, x, y; \theta) dtdxdy.$$

Here, S × [0,T] is the space–time window and *i* is an index for each flu activity occurring in region S and time interval [0, T]. In this study, *S* represents U.S., and *T* corresponds to 181 days of the 2013-2014 flu outbreak.

3.4 Forecasting

We built two point processes dealing with Twitter flu activities and clinical flu encounters in Cerner medical records. Table 5 summarizes the parameter estimates of the two ETAS models.

We used Kolmogorov-Smirnov (K-S) test to examine goodness-of-fit for each model separately. We summarize the results of the K-S tests in Table 6 demonstrating how well each model performs on the observed data. It contains the K–S test statistic value (D) corresponding to each model.

The results of the K-S tests indicate that an epidemic-type Hawkes process with a power law response function exhibits similar spatial and temporal characteristics of the flu outbreak in both Twitter and medical datasets. We subsequently used the R package "MPPA" (Rubin-Delanchy and Heard 2014) to examine the lag or lead dependence between the two point processes, and hence, to assess the spatio-temporal relationship between the actual flu cases and the social media activity.

In this study, with geo-tagged and time-stamped Twitter and clinical data at hand, we developed mashup applications to derive valuable insights from such diverse data sources. In doing so, Google's maps API libraries (Gesmann and de Castillo 2013; M Gesmann et al. 2013) were employed to identify several public locations from which a majority of tweets were originated over the course of time. In the next section, we will present and discuss the results of our temporal-spatio forecasting analysis.

4 Results

4.1 Temporal Analysis

The goal of temporal analysis was to examine whether online activities on Twitter reflect the occurrence of flu outbreak. Our assumption was that people post tweets about flu more often when the flu outbreak is imminent. Figures 3 and 4 respectively show histograms of the Cerner's flu encounters data and the Twitter flu activity data during the 2013-14 outbreak season. Figure 4 indicates that 'influenza' and 'flu' appear more frequently on Twitter as the flu moves through the population each winter. To examine the correlation between the clinical flu encounters and flu-related activities on Twitter during the flu outbreaks, rather than using the Pearson correlation, which is widely used in prior research and assumes a zero lag between online and real-world activities, we first used a time-series analysis approach to obtain the temporal cross-correlation between the two trends. In time series analysis, a temporal cross-correlation is a measure of similarity of two-time series as a function of the lag of one relative to the other. We found that clinical flu encounters lag 1 month behind online posts, and the number of unique users posting about flu per month is a good measure of the number of patients who visited a healthcare facility for ILI symptoms and were diagnosed with flu based on the Cerner data. Furthermore, we observed that the cross-correlation coefficient can be as high as 0.90 between the clinical flu encounters and Twitter flu activity. These results support that flu-related traffic on social media is closely related with actual flu outbreaks. This relation can be characterized by the theory of social exchange in which the discussions around flu has become somewhat of a phenomenon and that the utilization of social media to spread information about flu has changed the way people respond to the disease outbreak. It is pertinent to mention that this result does not imply that same individuals who tweet about their flu symptoms would visit a healthcare facility after 1 month. Rather, it simply points out to the lag between the trends that are observed in these two worlds. Figure 5 illustrates the cross-correlation test.

Table 6 The K-S goodness-of-fit test

K-S test values for the ETAS models	
Twitter (A)	Cerner (B)
D= 0.1336, *p* value = 7.84e-4*	D=0.2687, p value = 1.29e-2*

*Significant at 0.05

Table 5 Parameter value estimates

Process	μ	α	β	p	d	q
Twitter (A)	0.000853	0.182900	0.109900	0.909970	2.381454	1.115473
Cerner (B)	0.055419	0.936570	0.709437	1.874560	1.836540	1.306577

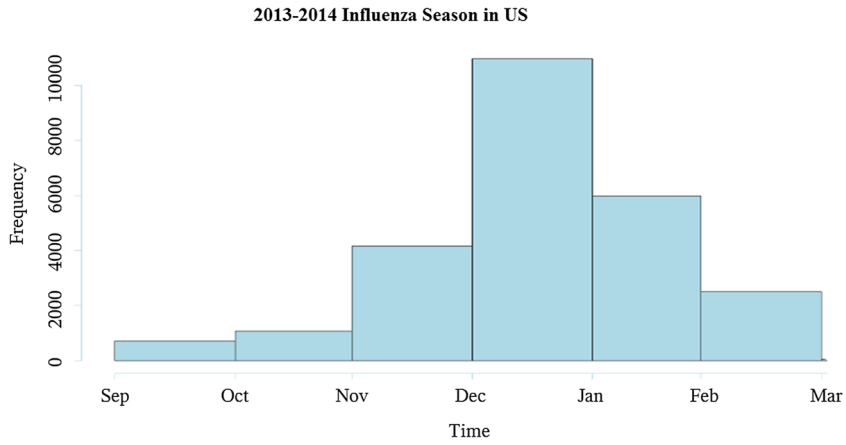

Fig. 3 Flu Activity in Cerner flu encounters data during the 2013-14 outbreak

4.2 Spatial Analysis

The goal of the spatial analysis was to track the geographic spread of influenza activities using information gathered from microblogging websites such as Twitter. Our Twitter dataset contains all the flu-related tweets that have conterminous US geographic coordinates. We excluded all tweets originated from outside of the US; and also ignored those users who generated flu tweets with invalid geo-location information. Figure 6 shows the tweet density for the US. The red "hotspots" indicate that a high number of people, normalized by zip code, tweeted about flu from that location. Guided by the critical mass theory, Fig. 6 presents a snapshot of critical mass achievement at locations across the US. The larger, red markers indicate locations that achieved critical mass of hotspots during the 2013-2014 outbreak. Overall, we can observe that people from different places in the US tweet about their flu, and a majority of these tweets originate from the eastern US; while the highest density of tweets in the western US are in California. Mid-eastern US has also very high density of flu tweets; however, people from mid-western US do not tweet as much about flu. Southern part of the US, such as Texas and Florida, also shows high a density of tweets. The density of tweets in other states is roughly even. It is important to acknowledge that normalization by Twitter subscriptions, while more appropriate, is not possible due to lack of such data. Therefore, normalization by population is an approximation.

Next, Moran's spatial autocorrelation (Moran 1950) was used to test the overall spatial randomness of flu activities within both Twitter and Cerner data at a 5% significance level. The Moran's I statistic is designed to analyze patterns occurring across the space and ranges from −1 to +1, with zero being the expected value for no spatial autocorrelation. This statistic is useful because it measures the spatial dependence between neighborhood property values compared to what would be expected to occur by chance. Moran's spatial index integrates space and spatial

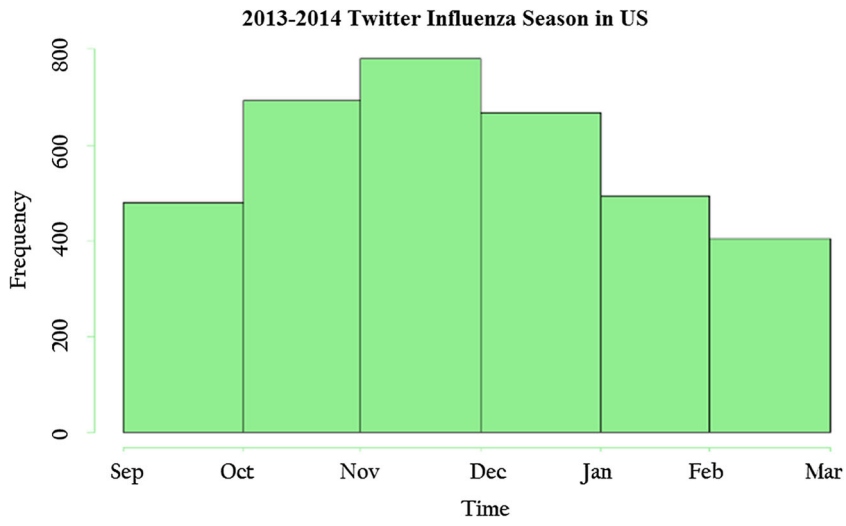

Fig. 4 Flu-related Twitter Activity during the 2013-14 outbreak

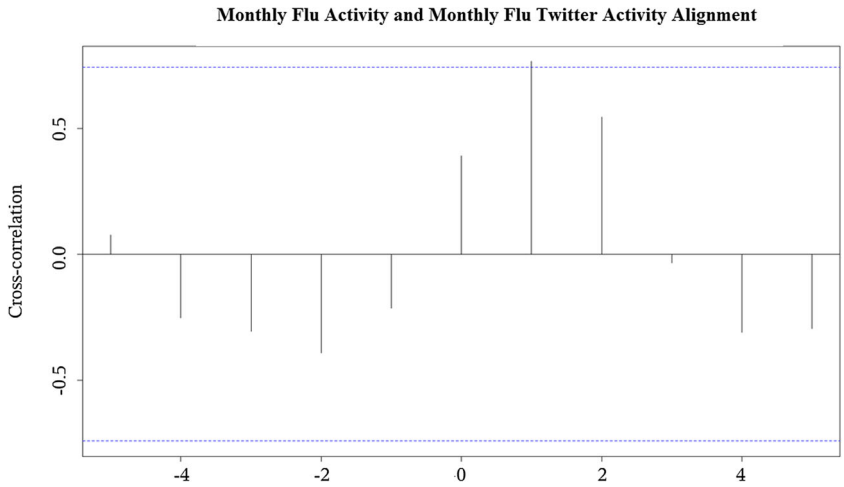

Fig. 5 Cross-Correlation between Twitter flu activity and clinical flu encounters

relationships directly into its mathematics through values called spatial weights. Weights are a measure of geographic proximity and reflect how connected two areas are. While there are many ways that can be used to specify spatial weights such as fixed distance, inverse distance, K-nearest neighbors, binary contiguity or some other criteria reflecting the alternative spatial relationship, we used an inverse distance function to calculate weights based on geographic distance. This is a reasonable model to use when one is interested in detecting contiguous areas with similar patterns in continuous data. In addition, no specific threshold distance or cutoff point is assumed and hence, the default setting ensures that each observation has at least one neighbor and will still have a minimal impact on far away areas. With our matrix of weights, we then calculated a single (global) Moran's I statistic to test for spatial autocorrelations in both Twitter and Cerner datasets. The results of the Moran's spatial autocorrelations are significant at p value <0.05 (see Table 4). These results prove the existence of a correlation between the geographical spread of flu within both datasets. For purpose of this analysis, we aggregated our Twitter and Cerner medical data to the state-level during the integration/fusion phase, so that both subsets are at the same geographic unit for our geo-spatial analysis. Therefore, the Moran's I spatial autocorrelation coefficients reported in Table 7 were measured at the state level.

In order to determine areas with a high risk of flu-related complications, i.e. hotspots, we identified public locations from which a majority of flu-related tweets were posted over the course of time. To find places from which people tweeted more frequently, we aggregated the data points by latitude and longitude and sorted them by number of tweets. Next, we normalized them on the basis of population at the zip code level in order to account for population density. Using Google Maps API tools, we were able to zoom in to identify and inspect these locations. For example, we found that Forest Park, New York,

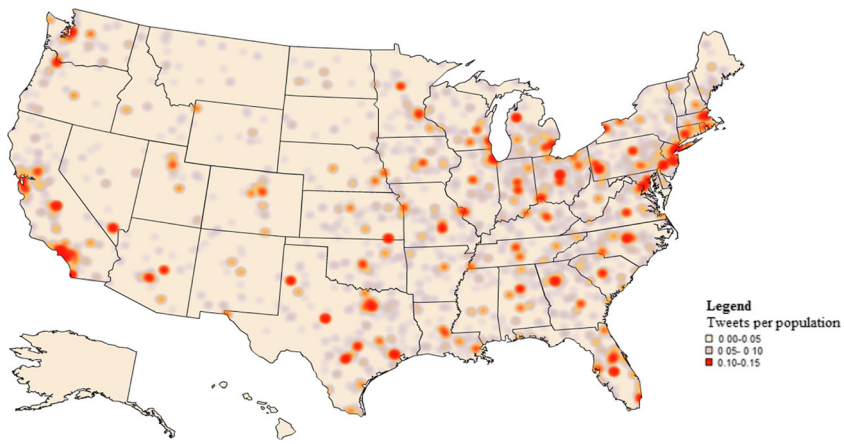

Fig. 6 GIS-based density map of the US flu surveillance

Table 7 Spatial Autocorrelation of Twitter and Cerner Flu datasets, U.S. 2013-2014, as measured by Moran's I Statistic

Moran's I values for the U.S. Flu Outbreak	
Cerner (C)	Twitter (T)
N_C= 50	N_T= 50
I_C=0.415*	I_T=0.669*

*Significant at 0.05; N: Sample size

I: Moran's I coefficient.

NY 10007 was one of the places from which many flu-related tweets originated. This place is a non-residential area with many shops and offices. Looking at the word network of the tweets originating from this place (as shown in Fig. 7) revealed that most of the tweets were from people who were affected by flu. There were no evidence of tweets advertising flu remedies and products to people, as we had filtered out such non-flu-symptom related tweets. More importantly, the word network displays flu symptoms that are mostly related to one another based on the thickness of the link between the nodes. Expectedly, being sick is the symptom that is central to the network, as it is generally mentioned when users tweet about flu.

The second location with the highest number of tweets represents Sound Walk, a public entertainment place in New York, US. As another example, the next location with the highest number of tweets is Disney Adventure Park in California. Interestingly, many people tweeted about their flu symptoms while involved in pastime.

All top ten locations in terms of the frequency of flu-related tweets were such public, non-residential places as parks, restaurants, hotels, universities and shopping stores (See Table 8). Regarding the number of people who visit these places every day, they play a significant role in the spread of the flu virus. Therefore, public health agencies can benefit from these findings by tracking and locating high-risk places, and highlighting "must-know" facts during the outbreaks. This information can help people who plan to visit such locations to take appropriate measures. Note that hotspots are meant to be places where many flu-related tweets originated.

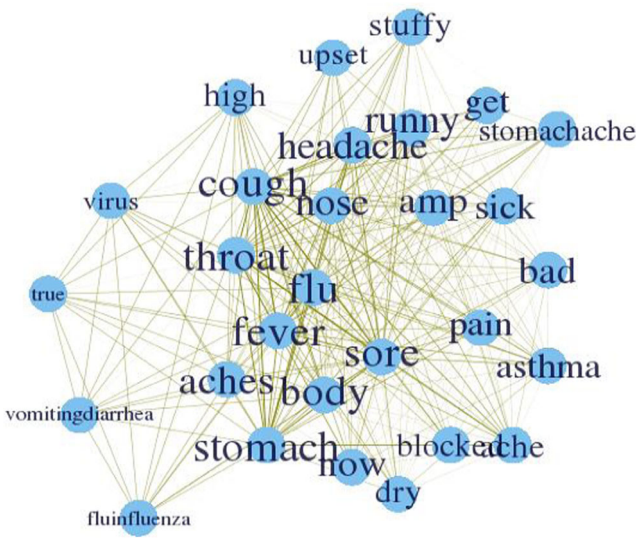

Fig. 7 Word network of the highest tweet activity place: Forest Park, New York, NY

It goes without saying that our primary goal is to examine if there is any association between social media and flu behavior at the aggregated (state) level. Once we found out that there is a strong correlation between the flu trends in both datasets, we dug into the Twitter dataset to identify suspicious flu-infected spots at the zip code level. We were not interested in testing the relationship between flu activity in both datasets at the zip code level and probably it would not make sense to perform such an analysis. Flu infected people would not visit a healthcare facility necessarily in the same zip code where they tweeted about their flu. For example, people tweeting about their flu in Disneyland, a university or an airport may not necessarily go to a healthcare facility at the same zip code to get the treatment. Even if we tested their relationship at the zip code level, it would not give us any useful insights.

4.3 Spatio-Temporal Analysis

As outlined in the method and data section, we proposed a new statistical measure to identify lag or lead correlations between events that contain both temporal and spatial components (at 5% significance level). Our measure uses a point processes approach that aims to the test lag or lead dependence between two spatio-temporal patterns, in our case, i.e. flu-related activities on Twitter (A) and clinical flu encounters in Cerner medical records (B) during the

Table 8 Top 10 locations with the largest number of tweets about flu

Location	Type
Forest Park, NY	Non-residential/Park
Sound Walk, NY	Entertainment
Disney Adventure Park, CA	Entertainment & Resort
Shopping mall, OH	Shopping, Dining and Hotel
Community College, KS	Campus
Niagara Falls, NY	Tourism
Disney World, FL	Entertainment
Airport, IL	Airport
Convention Center, TX	Non-residential
Public Library, CA	Central library

outbreak season. Using the Multiple Point Process Analysis (MPPA) approach (Rubin-Delanchy and Heard 2014), we developed a generalized likelihood ratio test statistics to examine the hypothesis that the relative proportion of events in A trigger the intensity of B. Under the null hypothesis, the events in B are assumed to follow process A. In the next subsection, we test this hypothesis empirically using the point processes we built in the model development section.

Statistical Inference and Hypothesis Testing In this section, we explain the Multiple Point Process Analysis approach used to examine the spatio-temporal relationship between the flu trends in social media and medical records within each region. Let process $A = \{a_1 < \ldots < a_m\}$, $m \geq 1$ represent the occurrence times of all the flu-related activities on Twitter and process $B = \{b_1 < \ldots < b_n\}$, $n \geq 0$, $b_1 \geq a_1$ be a set of event times corresponding to all the flu cases in Cerner medical records during the outbreak season (Note: both A and B are simulated using estimated ETAS parameters outlined in Table 6). Using the time transformation theorem (Daley and Vere-Jones 2007) (p.421), the intensity function of process B $\lambda_B(t,x,y)$ can be formulated as follows:

$$\lambda_B(t,x,y) = \begin{cases} \lambda_1 r(t,x,y) & t - a(t,x,y) \leq \tau \\ \lambda_2 r(t,x,y) & t - a(t,x,y) > \tau \end{cases} \qquad t \in [0, L), \tag{1}$$

where $a(t,x,y)$ is the most recent event in process A occurring prior to t, $\lambda_1 \geq \lambda_2 \geq 0$ are unknown parameters and r is a non-negative measurable intensity function satisfying $\int_0^L r(v)dv = 1$. Given eq. (1), the following hypothesis test is used to examine whether events in A cause an increase in the intensity of B (triggering behavior both in time and space):

$$H_0 : \lambda_1 = \lambda_2 \quad H_1 : \lambda_1 > \lambda_2$$

where the test statistic is the generalized likelihood ratio

$$T = \frac{\sup\{L(B;\tau,\lambda_1,\lambda_2) : \tau > 0, \lambda_1 > \lambda_2 \geq 0\}}{\sup\{L(B;\tau,\lambda_1,\lambda_2) : \tau > 0, \lambda_1 = \lambda_2 \geq 0\}}$$

Here L is the likelihood of B under model (1). Under the null hypothesis, A and B are functionally independent, whereas under the alternative hypothesis, the relative proportion of events in A is higher than can be explained by r alone. In other words, events in A trigger the intensity of B. We found that the statistical test for causal and mutual dependence between A and B (B events are caused/triggered by A) become significant at p value <0.05 for all regions. It supports that flu-related activities on social media can lead to an early detection of influenza outbreaks within each region. We summarize the results of the dependence test in Table 9.

Table 9 The results of interaction or dependence analysis

Dependence test
$T = 0.1105$, p value <2.2e-16*

*Significant at 0.05

In summary, the findings of this study have the potential to create awareness about general patterns of public behavior and social interactions/buzz that indicate the emergence of certain flu pandemics. Our results can also increase the accuracy of locating flu hotspots. Moderating influenza levels in these areas leads to the greatest spillover effects to nearby areas. Our findings are in line with theories of critical mass, social exchange, and attachment motivation in that how spatio-temporal dynamics of information exchange affect the decisions of both patients and health professionals. Above all, this paper aimed to pave the ground for such research work towards the use and implementation of big data technologies for analytical purposes.

This study provides theoretical, empirical, and methodological advances that help policymakers use location-based social media platforms as a source of spatio-temporal information to detect the spread of various diseases. The point process approach outlined in this paper allows us to quantify spatial and temporal dependencies among events that arise as a result of cascading behaviors. It computes cross-correlation of two point process directly without any binning of the input. Our approach takes two sets of data points as input, that are at the lowest possible level of granularity (i.e. the timestamp and geo-location level) and automatically measures spatio-temporal associations between the two observed patterns at the lowest level.

5 Discussion and Limitations

With social media continually generating vast amounts of data, big data analytics is coming to the forefront of ways to turn this data into useful information for uses ranging from health promotion and education to crisis management systems and public health surveillance. In this paper, we asserted that location-based social media, as a big health data source, can be leveraged as a feasible geo-specific monitoring tool to provide strategic insights for planning, execution and measurement of effective and efficient public health interventions. We showed that by integrating two sets of flu activity data, i.e., Twitter posts and real encounters, flu outbreaks can be nowcasted and consequently, better decisions can be made in managing such pandemics. The results of this study showed that online posts add a great value to the location-based nowcast of flu outbreaks. Combining and analyzing locations of such tweets will allow healthcare professionals to locate,

with much greater accuracy, where flu pandemics originate and implement behavioral interventions if needed. This enables the public to take safety precautions when traveling to flu ridden locations. Overall, these findings confirm that the current flu (or any infectious disease or pandemic) surveillance systems can be improved with the integration of real time social media posts. This positively impacts the general public and pinpoints specific geographic locations that are in need of greater resources to fight off flu during an outbreak. Even though big data has been shown to give a better idea of real-time flu trends compared to traditional data sources, it is considered to be more of an integration tool than a replacement to the existing disease surveillance systems. It is important that the right data is analyzed and processed when analyzing such large stores of data, which allows for informed decision making and more timely actionable intelligence for a prompt reaction from healthcare organizations during an emergency or outbreak. In this effort, the largest challenges appear to be legal, political and economic obstacles (Sane and Edelstein 2015).

We anticipate that our findings will inform clinical practice and public health policy. Location-enabled social media platforms have the potential to provide clinicians with real-time and geo-located data about contagious diseases, such as the seasonal influenza. This system will enable clinicians to plan interventions and proactive actions early enough to minimize an immense burden a flu outbreak can trigger. In addition, the Twitter Influenza Surveillance system increases the ability of state and local health officials to identify most flu-infected places in order to respond expeditiously and take appropriate actions. Last, but not the least, a similar system, based on a mobile app, can alert individuals about the suspicious locations they should avoid during those periods. Together, the findings of our study provide a proof-of-concept for the use of geo-located social media platforms to support public health initiatives at individual, local and national levels targeted at quickly discovering trends of contagious diseases before they escalate into epidemics. Our study provides theoretical, empirical, and methodological advances that can help researchers and practitioners use location-based social media platforms as a source of spatio-temporal information to detect the early spread of infectious diseases.

Our study is not without limitations. First, social media as sources for data do not contain detailed geolocation information for in-depth geographical analysis, because not all users have their GPS enabled or declare their location in their social media profiles. Without knowing the location of all users, spatial analysis of social media may be biased (Tsou 2015). As such, it is critical to use various techniques to identify and correct for such biases in data. Second, there still are sources of noise in social media, including robots, and non-relevant conversations. More advanced classification techniques or machine learning approaches should be used for deeper content analysis of social media posts. In addition, our analysis can be extended to study the epidemic spread of flu within different subpopulations by leveraging socio-economic and demographics data, resulting in a more targeted approach, and therefore, more effective course of detection and prevention. The point processes model outlined in this research is capable of incorporating the observed temporal patterns, specifying the spatial and temporal dependencies and predicting future outbreaks using the background processes of flu in both Twitter and medical datasets. Therefore, it could be a useful future research direction to build a mutual-exciting version of our proposed Hawkes processes (Hawkes 1971) that may predict the flu occurrences more accurately. Finally, it is important to acknowledge the Modifiable Areal Unit Problem (MAUP) as it pertains to this study. The fact that the data are aggregated into larger spatial units such as zip codes, census tracts, counties, or states may affect how the data are interpreted (Fotheringham and Wong 1991; O'Sullivan and Unwin 2014). The present study was limited in its ability to examine the modifiable areal unit problem as the scale and selection of the geographic unit of analysis was constrained by the format of the data available in the Cerner data warehouse.

6 Conclusions

Nowcasting with social media has become a major topic of interest to researchers and policy makers, assisting them in understanding public opinion and trends and in forecasting early detection of disease outbreaks, thus allowing timely response and intervention, which reduces the impact of the outbreak on public health. In this study, we demonstrated location-based nowcast of flu outbreaks using social media data. We used a big data analytics approach to integrate two sets of flu data. One was obtained from flu-related tweets and the other was the Cerner's clinical encounters. We analyzed extensive datasets consisting of daily activities of hundreds of thousands of people across the US from both Twitter and Cerner clinical records. We found that incorporating such dynamic information on public behavior could be a valuable addition to bolster the capacity of existing models to explain and nowcast the flu activity. We used a series of spatio-temporal analyses to describe the early spread of flu. In the temporal analysis, we examined whether Twitter data could be used for the nowcasting of influenza outbreaks. We tracked and compared clinical flu encounters and flu-related activities on Twitter during the outbreaks. In the spatial analysis, by the use of mashup applications, we mapped flu outbreaks to the geo-spatial property of Twitter data to identify influenza hotspots.

Drawing on the lead-lag theory in time series, we demonstrated the application of point processes as an integrated framework to study lag-lead relationship and spatio-temporal

cross-correlation in a location-based social network context. We examined not only the co-incidence of flu activity on Twitter and clinical health records, but also whether one preceded the other. A series of models were developed to examine the spatio-temporal relationship between the flu trends in social media and medical records. Our results supported that flu-related traffic on social media is closely related with actual flu outbreaks. We found that clinical flu encounters lag one month behind online posts. We observed that the online flu conversation tends to spike a month earlier than actual flu cases. Also, we identified several public locations from which a majority of posts originated. These findings show that location-based social media platforms can assist healthcare policymakers by providing spatio-temporal information for disease surveillance systems. Our result shows the integration of social media and medical records data, equipped with spatial big data tools and supported by geographic information systems, can facilitate early detection and rapid development of behavioral interventions that can assist public health agencies to control and prevent epidemics of such infectious diseases. Such efforts can be extended to a personalized location-based application and adopted by individuals to inform locations that are infected during those periods.

Social media could add value to existing disease surveillance systems in several ways. First and foremost, it improves the timeliness, temporal and spatial resolution of surveillance information. It also expands the surveillance coverage by adding surveillance to places with no existing systems where scarcity of resources or other constraints limit the availability of direct clinical or laboratory data. Furthermore, it measures aspects of the spreading process (i.e. behavior, perception, dynamics) that may not be captured by traditional surveillance. For example, our study illustrated that flu-related activities on social media can be used to identify high-risk places for flu infection, which cannot be otherwise detected via the existing physician and laboratory surveillance systems. There is no doubt that enhancing the surveillance systems with real-time visibility brought by social media can give public health officials the insight and details needed to operate quickly, accurately, and more effectively than ever before. Public health agencies can benefit from our analyses by tracking and locating high-risk places in real-time, and by developing more effective behavioral intervention approaches to improve the inspection process in order to minimize the spread of diseases in the affected areas.

In summary, the utility of social media as a behavioral intervention platform has the potential for profound impact. Examples include the US Centers for Disease Control and Prevention (CDC) Twitter feeds and the Skin Cancer Foundation Facebook page where bidirectional communication channels exist between patients and health care providers. Such platforms are of great value to public health agencies when the purpose of the intervention is to be able to generate conversations, answer participant questions, alter patients' behavior and attitudes and/or provide behavioral counseling in the in the face of a disease outbreak (Pagoto et al. 2016).

Acknowledgments This study was conducted with the data provided by, and the support from, the Center for Health Systems Innovation (CHSI) at Oklahoma State University (OSU) and the Cerner Corporation. The contents of this work are solely the responsibility of the authors and do not necessarily represent the official views of CHSI, OSU or the Cerner Corporation. We also want to thank the anonymous reviewers and the associate editor for very thoughtful comments on the paper.

References

Al-garadi, M. A., Khan, M. S., Varathan, K. D., Mujtaba, G., & Al-Kabsi, A. M. (2016). Using online social networks to track a pandemic: a systematic review. *Journal of Biomedical Informatics, 62*, 1–11.

Allen, C., Tsou, M.-H., Aslam, A., Nagel, A., & Gawron, J.-M. (2016). Applying GIS and machine learning methods to twitter data for multiscale surveillance of influenza. *PLoS One, 11*(7), e0157734.

Amorós, R., Conesa, D., Martinez-Beneito, M. A., & López-Quılez, A. (2015). Statistical methods for detecting the onset of influenza outbreaks: A review. *REVSTAT–Statistical Journal, 13*(1), 41–62.

Anselin, L. (1989). What is special about spatial data? Alternative Perspectives on Spatial Data Analysis (89-4).

Anselin, L. (2013). *Spatial econometrics: methods and models* (Vol. 4). Berlin: Springer Science & Business Media.

Aslam, A. A., Tsou, M.-H., Spitzberg, B. H., An, L., Gawron, J. M., Gupta, D. K., ... Yang, J.-A. (2014). The reliability of tweets as a supplementary method of seasonal influenza surveillance. *Journal of Medical Internet Research, 16*(11), e250.

Brillinger, D. R., Bryant, H. L., & Segundo, J. P. (1976). Identification of synaptic interactions. *Biological Cybernetics, 22*(4), 213–228.

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through twitter: an analysis of the 2012-2013 influenza epidemic. *PLoS One, 8*(12), e83672.

Chen, Y. (2015). A new methodology of spatial cross-correlation analysis. *PLoS One, 10*(5), e0126158.

Chen, Y.-D., Brown, S. A., Hu, P. J.-H., King, C.-C., & Chen, H. (2011). Managing emerging infectious diseases with information systems: reconceptualizing outbreak management through the lens of loose coupling. *Information Systems Research, 22*(3), 447–468.

Chorianopoulos, K., & Talvis, K. (2016). Flutrack.org: open-source and linked data for epidemiology. *Health Informatics Journal, 22*(4), 962–974.

Congdon, P. (2005). *Bayesian models for categorical data*. Hoboken: John Wiley & Sons.

Copeland, P., Romano, R., Zhang, T., Hecht, G., Zigmond, D., & Stefansen, C. (2013). Google disease trends: an update. *Nature, 457*, 1012–1014.

Corberán-Vallet, A., & Lawson, A. B. (2014). Prospective analysis of infectious disease surveillance data using syndromic information. *Statistical Methods in Medical Research, 23*(6), 572–590.

Daley, D. J., & Vere-Jones, D. (2007). *An introduction to the theory of point processes: Volume II: General theory and structure*. Berlin: Springer Science & Business Media.

Davidson, M. W., Haim, D. A., & Radin, J. M. (2015). Using networks to combine "big data" and traditional surveillance to improve influenza predictions. *Scientific Reports, 5*, 8154.

Dewan, S., & Ramaprasad, J. (2014). Social media, traditional media, and music sales. *MIS Quarterly, 38*(1), 101–122.

Dewan, S., & Ramprasad, J. (2009). Chicken and egg? Interplay between music blog buzz and album sales. *PACIS 2009 proceedings*, p. 87.

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter?—an empirical investigation of panel data. *Decision Support Systems, 45*(4), 1007–1016.

Dukic, V., Lopes, H. F., & Polson, N. G. (2012). Tracking epidemics with Google flu trends data and a state-space SEIR model. *Journal of the American Statistical Association, 107*(500), 1410–1426.

Fang, Z.-H., & Chen, C. C. (2016). A novel trend surveillance system using the information from web search engines. *Decision Support Systems, 88*, 85–97.

Fotheringham, A. S., & Wong, D. W. (1991). The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A, 23*(7), 1025–1044.

Gesmann, M., & de Castillo, D. (2013) googleVis: Using the Google Chart Tools with R.

Gesmann, M., de Castillo, D., & Cheng, J. (2013). googleVis: Interface between R and the Google Chart Tools. R package version 0.4, 2.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature, 457*(7232), 1012–1014.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature, 453*(7196), 779–782.

Granger, C. W. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society, 37*(3), 424–438.

Griffin, B. A., Jain, A. K., Davies-Cole, J., Glymph, C., Lum, G., Washington, S. C., & Stoto, M. A. (2009). Early detection of influenza outbreaks using the DC Department of Health's syndromic surveillance system. *BMC Public Health, 9*(1), 483.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika, 58*(1), 83–90. https://doi.org/10.2307/2334319.

Homans, G. C. (1958). Social behavior as exchange. *American Journal of Sociology*, 597–606.

Lamb, A., Paul, M. J., & Dredze, M. (2013). Separating Fact from Fear: Tracking Flu Infections on Twitter. Paper presented at the HLT-NAACL.

Lampos, V., Miller, A. C., Crossan, S., & Stefansen, C. (2015). Advances in nowcasting influenza-like illness rates using search query logs. *Scientific Reports, 5*, 12760.

Lara Yejas, O. D., Weiqiang, Z., & Pannu, A. (2014). Big R: Large-Scale Analytics on Hadoop Using R. Paper presented at the Big Data (BigData Congress), 2014 IEEE International Congress on.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science, 343*(6176), 1203–1205.

Louis, C. S., & Zorlu, G. (2012). Can Twitter predict disease outbreaks? *BMJ: British Medical Journal (Online), 344*(7861), 24–25.

Lymperopoulos, I. N., & Ioannou, G. D. (2015). Online social contagion modeling through the dynamics of integrate-and-fire neurons. *Information Sciences, 320*, 26–61.

Ma, J., Zeng, D., & Chen, H. (2006). Spatial-temporal cross-correlation analysis: a new measure and a case study in infectious disease informatics. Paper presented at the International Conference on Intelligence and Security Informatics.

Magruder, S. (2003). Evaluation of over-the-counter pharmaceutical sales as a possible early warning indicator of human disease. *Johns Hopkins APL Technical Digest, 24*(4), 349–353.

Milinovich, G. J., Williams, G. M., Clements, A. C. A., & Hu, W. (2014). Internet-based surveillance systems for monitoring emerging infectious diseases. *The Lancet Infectious Diseases, 14*(2), 160–168. https://doi.org/10.1016/S1473-3099(13)70244-5.

Mohler, G. (2014). Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting, 30*(3), 491–497.

Molinari, N.-A. M., Ortega-Sanchez, I. R., Messonnier, M. L., Thompson, W. W., Wortley, P. M., Weintraub, E., & Bridges, C. B. (2007). The annual impact of seasonal influenza in the US: measuring disease burden and costs. *Vaccine, 25*(27), 5086–5096. https://doi.org/10.1016/j.vaccine.2007.03.046.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika, 37*(1/2), 17-23.

Nguyen, B. V., Burstein, F., & Fisher, J. (2015). Improving service of online health information provision: a case of usage-driven design for health information portals. *Information Systems Frontiers, 17*(3), 493–511.

Nunes, B., Natário, I., & Lucília Carvalho, M. (2013). Nowcasting influenza epidemics using non-homogeneous hidden Markov models. *Statistics in Medicine, 32*(15), 2643–2660.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association, 83*(401), 9–27.

Oliver, P., Marwell, G., & Teixeira, R. (1985). A theory of the critical mass. I. Interdependence, group heterogeneity, and the production of collective action. *American Journal of Sociology, 91*(3), 522-556.

O'Sullivan, D., & Unwin, D. (2014). *Geographic information analysis*. Hoboken: John Wiley & Sons.

Pagoto, S., Waring, M. E., May, C. N., Ding, E. Y., Kunz, W. H., Hayes, R., & Oleski, J. L. (2016). Adapting behavioral interventions for social media delivery. *Journal of medical Internet research, 18*(1), e24. https://doi.org/10.2196/jmir.5086.

Patwardhan, A., & Bilkovski, R. (2012). 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, 7*(8), e43611.

Pick, J. B., Sarkar, A., & Johnson, J. (2015). United States digital divide: state level analysis of spatial clustering and multivariate determinants of ICT utilization. *Socio-Economic Planning Sciences, 49*, 16–32.

Prati, G., Pietrantoni, L., & Zani, B. (2011). A social-cognitive model of pandemic influenza H1N1 risk perception and recommended behaviors in Italy. *Risk Analysis, 31*(4), 645–656.

Richards, C. L., Iademarco, M. F., & Anderson, T. C. (2014). A new strategy for public health surveillance at CDC: improving national surveillance activities and outcomes. *Public Health Reports, 129*(6), 472–476.

Rubin-Delanchy, P., & Heard, N. A. (2014). A test for dependence between two point processes on the real line. *arXiv preprint arXiv: 1408.3845*.

Rudra, K., Sharma, A., Ganguly, N., & Imran, M. (2018). Classifying and summarizing information from microblogs during epidemics. *Information Systems Frontiers*, 1-16. https://doi.org/10.1007/s10796-018-9844-9.

Sane, J., & Edelstein, M. (2015) Overcoming barriers to data sharing in public health. A global perspective. London: Chatham House.

Santillana, M., Nguyen, A. T., Dredze, M., Paul, M. J., Nsoesie, E. O., & Brownstein, J. S. (2015). Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Computational Biology, 11*(10), e1004513.

Santillana, M., Nguyen, A. T., Louie, T., Zink, A., Gray, J., Sung, I., & Brownstein, J. S. (2016). Cloud-based Electronic Health Records for Real-time, Region-specific Influenza Surveillance. *Scientific Reports, 6*, 25732.

Sebastiani, P., Mandl, K. D., Szolovits, P., Kohane, I. S., & Ramoni, M. F. (2006). A Bayesian dynamic model for influenza surveillance. *Statistics in Medicine, 25*(11), 1803–1816.

Shi, Z., Rui, H., & Whinston, A. B. (2014). Content sharing in a social broadcasting environment: evidence from twitter. *MIS Quarterly, 38*(1), 123–142. https://doi.org/10.25300/misq/2014/38.1.06.

Signorini, A., Segre, A. M., & Polgreen, P. M. (2011). The use of twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. *PLoS One, 6*(5), e19467.

Simonsen, L., Gog, J. R., Olson, D., & Viboud, C. (2016). Infectious disease surveillance in the big data era: towards faster and locally relevant systems. *The Journal of Infectious Diseases, 214*(suppl_4), S380–S385.

Talvis, K., Chorianopoulos, K., & Kermanidis, K. L. (2014). Real-time monitoring of flu epidemics through linguistic and statistical analysis of Twitter messages. Paper presented at the Semantic and Social Media Adaptation and Personalization (SMAP), 2014 9th International Workshop on.

Toole, J. L., Eagle, N., & Plotkin, J. B. (2011). Spatiotemporal correlations in criminal offense records. *ACM Transactions on Intelligent Systems and Technology, 2*(4), 1–18. https://doi.org/10.1145/1989734.1989742.

Tsou, M.-H. (2015). Research challenges and opportunities in mapping social media and big data. *Cartography and Geographic Information Science, 42*(sup1), 70–74.

Vandendijck, Y., Faes, C., & Hens, N. (2013). Eight years of the great influenza survey to monitor influenza-like illness in Flanders. *PLoS One, 8*(5), e64156.

von Alan, R. H., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly, 28*(1), 75–105.

Wagner, M., Lampos, V., Cox, I. J., & Pebody, R. (2018). The added value of online user-generated content in traditional methods for influenza surveillance. *Scientific Reports, 8*(1), 13963. https://doi.org/10.1038/s41598-018-32029-6.

Wang, D.-H., Suo, Y.-Y., Yu, X.-W., & Lei, M. (2013). Price–volume cross-correlation analysis of CSI300 index futures. *Physica A: Statistical Mechanics and its Applications, 392*(5), 1172–1179.

Wilson, K., & Brownstein, J. S. (2009). Early detection of disease outbreaks using the internet. *Canadian Medical Association Journal, 180*(8), 829–831.

Young, S. D., Rivers, C., & Lewis, B. (2014). Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. *Preventive Medicine, 63*, 112–115.

Amir Hassan Zadeh is an associate professor of MIS at Wright State University. He received his PhD in Management Information Systems from Oklahoma State University, OK, US. His research and teaching interests are in enterprise systems and applications, business analytics, data and text mining, and big data applications. His research works have appeared in journals such as Decision Support Systems, Production Planning & Control, Journal of Cases on Information Technology, International Journal of Advanced Manufacturing Technology among others. He serves as the marketing and communication co-chair of ICIS Decision Support and Analytics (SIGDSA) symposium.

Hamed M. Zolbanin is an assistant professor of information systems and the director of business analytics in Miller College of Business at Ball State University. He had several years of professional experience as an IT engineer prior to receiving his Ph.D. in Management Science and Information Systems from Oklahoma State University. His research has been published in such journal as Decision Support Systems and Journal of Business Research. His main research interests are healthcare analytics, sharing economy, and digital entrepreneurship.

Ramesh Sharda is the Vice Dean for Research and Graduate Programs, Watson/ConocoPhillips Chair and a Regents Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University. He has coauthored two textbooks (Business Intelligence, Analytics, and Data Science: A Managerial Perspective, 4th Edition, Pearson and Business Intelligence and Analytics: Systems for Decision Support, 10th edition, Pearson). His research has been published in major journals in management science and information systems including Management Science, Operations Research, Information Systems Research, Decision Support Systems, Decision Sciences Journal, EJIS, JMIS, Interfaces, INFORMS Journal on Computing, ACM Data Base and many others. He is a member of the editorial boards of journals such as the Decision Support Systems, Decision Sciences, and Information Systems Frontiers. He also serves as the Faculty Director of Teradata University Network. He received the 2013 INFORMS Computing Society HG Lifetime Service Award, and was inducted into Oklahoma Higher Education Hall of Fame in 2016. Ramesh is a fellow of INFORMS.

Dr. Dursun Delen is the holder of Spears Endowed Chair in Business Administration, Patterson Family Endowed Chair in Business Analytics, Director of Research for the Center for Health Systems Innovation, and Regents Professor of Management Science and Information Systems in the Spears School of Business at Oklahoma State University (OSU). Dr. Delen has over 30 years of experience in analytics both as a business consultant and university professor. Prior to his academic tenure, he worked for a privately-owned research and consultancy company as a research scientist for five years, during which he led a number of decision support, information systems and advanced analytics related research projects funded by industry and federal agencies including DoD, NASA, NIST and DoE. Dr. Delen has published more than 140 peerreviewed articles and eight books/textbooks. He is often invited to national and international conferences for keynote addresses, and companies for consultancy engagements on topics related to business analytics, data/text mining, and knowledge management. He regularly serves as chair for tracks and mini-tracks at various business analytics and information systems conferences. Currently, he is serving on more than a dozen journal editorial boards as editor-in-chief, senior editor, associate editor, and editorial board member. He is the recipient of several research and teaching awards including the prestigious Fulbright scholar, regents distinguished teacher and researcher, and Big Data mentor awards.