Social Media As A Public Health Surveillance Tool: Evidence And Prospects
Executive Summary

Social media has become an increasingly relevant tool for public health surveillance, with organizations such as Sickweather using social media to detect, monitor and predict epidemic trends to facilitate preparedness and rapid response. This work is especially pertinent in light of the emergence of globalized and growing epidemics for which there is no known cure. This project reviews the literature on the use of social media for public health surveillance. A total of 103 articles were reviewed to this end.

The studies reviewed presented considerable evidence on the effectiveness of social media as a public health surveillance tool with benefits such as accuracy, timeliness and cost-effectiveness in tracking population health sentiments, behaviours, outcomes and emergencies. Secondly, social media based public health surveillance provides an avenue for collaboration between various public health stakeholders and the dissemination of information between them. Social media based public health surveillance allows for the democratization of public health knowledge. By providing geo-coded information and spatiotemporal insights, social media allows for more granular awareness of population health issues.

Current challenges in the use of social media for public health surveillance concern the need for the elimination of noise from surveillance data as this interferes with effective surveillance. In addition, the use of public information calls into consideration ethical issues regarding privacy. There is need for more scientific research on the validity of surveillance data and mining. There is also need to build infrastructural capacity in health institutions for big data. If explored, these opportunities will yield more advanced insights on how to optimize the use of social media for public health surveillance.

The future of social media based public health surveillance lies in the use of social media to not only identify health outcomes but as a possible avenue for forecasting population health trends and for delivering interventions. While researchers do not view social media as a stand-alone means for public health surveillance in light of concerns for the scientific veracity of surveillance procedures, there is strong interest in the use of social media to supplement traditional surveillance infrastructure.

Therefore, while there is scientific basis for the use of social media based surveillance systems, future studies will need to move beyond observational analyses to testing the procedures for mining crowd sourced data as well as the validity of social media for forecasting public health trends. Furthermore, future research is needed on how to effectively integrate social media sources with clinical surveillance systems. This will ensure that health information systems are both scientifically valid while flexible enough to respond to health emergencies and outcomes.
Table of Contents

- Executive Summary
- Introduction
- Methodology
- Flowchart
- Findings
  - Cost-effectiveness
  - Supporting the spatio-temporal analyses of public health information
  - Enabling information dissemination
  - Tracking population health outcomes and preparing for disasters or outbreaks
  - Tracking public health attitudes and behaviours
- Conclusion
  - The need for stronger spatiotemporal forecasting models
  - The need for infrastructural capacity for big data
  - The potential of social media as an intervention site
  - The need to strengthen data mining and data validation
  - Ethical issues
  - Integrating social media data with traditional public health surveillance infrastructure
  - Strengthening lexicon development and models for text mining and classification
- References
Introduction

Public health surveillance systems enable the monitoring of public health trends and subsequent decision-making (Gilbert et al., 2016, Teutsch et al., 2000). Key features of public health surveillance systems are that they are continual, systemic, collect, analyze and interpret health data (World Health Organization 2017, Gilbert et al., 2016, Teutsch et al., 2000). Surveillance can provide early warning for impending disaster, record keeping for the purpose of public health monitoring and evaluation and an evidence base for epidemiology, as well as for decision making (World Health Organization, 2017).

While traditional clinical databases were the only surveillance systems available in the past, technological advancements are increasingly shaping the future of public health (Gilbert et al., 2016), and have led to the emergence of big data and more participatory means and norms of communication. Such digital technologies include cell phones, social media, search engines, mailing lists, electronic health records, participatory surveillance systems, blogs, news aggregators, media reports, social networks, wikis, visualization tools, mapping technologies, RSS feeds, discussion forums, web sites, news outlets, to mention but a few (Bansal et al., 2016, Brownstein et al., 2009, Brownstein et al., 2008, Chew et al., 2010).

These digital means of surveillance have been said to have advantages over more traditional means due to their timeliness and cost-effectiveness in monitoring public health outcomes (Gittelman et al., 2010, Kass-Hout et al. 2013). However, the increasing relevance of social media in public health surveillance has also raised concerns that pertain to the validity of these surveillance methods and their role in relation to traditional public health infrastructure (Kass-Hout et al., 2013, Moorhead et al., 2013, Kostkova et al., 2013, Brownstein et al., 2009). The focus of this project was therefore to conduct a systematic review of the literature specifically on the use of social media platforms such as Twitter and Facebook in public health surveillance. The goal of this was to yield an understanding of the current developments in social media public health surveillance, the evidence base for this means of surveillance, best practices in collecting and analyzing social media data for public health, as well as opportunities for further research in this field.

Methodology

In order to understand the current evidence base on the use of social media sites for public health surveillance, a systematic literature review was conducted on the Google Scholar database. Articles published from 2007 and onwards were included in the search, with the following keywords guiding the search: “social media for public health surveillance”, “social media and ILI reporting”, “social media and flu reporting”, “social media and disease reporting”, “social media and health behaviour reporting”, “social media and allergy reporting”, “predictive value of social media for public health”.

Figure 1 below shows a flowchart for the identification of documents to be included in the literature review. The resulting 103 articles were reviewed to explore contemporary uses of social media for public health surveillances, best practices in the use of social media for public health surveillance, strengths and limitations in the use of social media for public health surveillance and opportunities for further research.
METHODOLOGY: FIGURE 1 : FLOWCHART FOR LITERATURE REVIEW

Initial Search

Number of Articles Found: 107

Database: Google Scholar
Search Criteria: “social media for public health surveillance”, “social media and ILI reporting”, “social media and flu reporting”, “social media and disease reporting”, “social media and health behaviour reporting”, “social media and allergy reporting”, “predictive value of social media for public health”. Inclusion criteria: articles published after 2007.

3 articles that did not discuss social media surveillance were removed due to irrelevance to the focus of the review.

1 citation for an article with a future publication date removed.

Articles in Final Short List: 103

“OSN data can act as supplementary indicator to gauge influenza levels within the population and helps to discover flu trends ahead of CDC.” Achrekar, 2013
Findings

Below, evidence for the use of social media for public surveillance is reviewed including future directions and best practices for this field of practice. In addition, current limitations in the use of social media for public health surveillance and opportunities for further research are presented.

Cost-effectiveness: Several studies highlighted the advantages of using the internet for public health surveillance. One of these advantages was the cost-effectiveness of online sources given that current means for tracking health epidemics such as hospital surveys are quite expensive (Culotta, 2010; Gittelman et al., 2010).

Supporting the spatio-temporal analyses of public health information: Digital and informal means for public health surveillance allow for more granular analyses of epidemiological information (Bansal et al., 2016). Social media platforms such as twitter allow for localized and characterized information, as was the case during the Boston marathon bombings (Cassa et al., 2013). As a result, these digital surveillance platforms allow for rapid and focused detection of diseases given the much finer temporal and spatial resolution of disease monitoring information that they provide (Chunara et al., 2013).

In addition to providing more detailed geographical information, several studies pointed to the time advantage that social media has over traditional surveillance systems. Traditional surveillance systems are retrospective and tend to give a true yet retrospective and delayed assessment of epidemics (Chakraborty et al., 2014; Achrekar et al., 2013). Social media present affordable and early ways to detect and respond to outbreaks before are picked up by formal sources (Bernado et al., 2013; Brownstein et al., 2009; Salathe et al., 2013; Paul et al., 2014; Chew et al., 2010). Informal sources such as twitter, new media reports and digital surveillance tools were able to pick up cholera-related information up to two weeks earlier than traditional sources during the cholera outbreak in Haiti in 2010 (Chunara et al., 2012). Twitter also picked up news of the Boston Marathon Bombings before traditional sources (Cassa et al., 2013). While facilitating situational awareness, location based social media data and the rich data they capture allow the spatiotemporal dynamics of diseases to be captured at multiple scales (Padmanabhan et al., 2013).

Enabling information dissemination: Social media also allows for the dissemination of public health information by health organizations and practitioners (Keller et al., 2014; Chunara et al., 2012; Chew et al., 2010). This involves the use of platforms like Twitter for communicating with the public on surveillance and interventions (Kass-hout et al., 2013) and for sharing resources (Terry, 2009; Chew et al, 2010; Capurro et al, 2014). In addition, social media platforms such as Facebook forums allow members of the public avenues for sharing their experiences, receiving feedback and asking questions to individuals with whom they share health conditions such as diabetes (Greene et al., 2011).

Additionally, social media surveillance allows for the democratization of public health knowledge thus preventing government suppression of health emergencies (Brownstein et al., 2009). By democratizing public health information, social media connects health experts and facilitates the exchange of knowledge, the public’s access to health information, and the opportunities for peer, social and/or emotional support (Moorhead et al., 2013).
Tracking population health outcomes and preparing for disasters or outbreaks: Given the rising emergence of infectious diseases for which they are no cures yet, social media provides an opportunity for real time reporting on these diseases (Hartley, 2014). Clinically validated surveillance systems are retrospective given that they need to be compiled, thus leading to a delayed awareness of public health outcomes (Anoshe et al., 2014). Social media surveillance can thus allow for public health officials to detect, monitor, manage and respond to epidemics and disasters (St. Louis et al., 2012, Veil et al., 2011, Fung et al., 2013, Kumar et al., 2015, Achrekar et al., 2013), track health outcomes over time, track disease flows, monitor social networks (Schmidt, 2012, Stoove et al., 2012, Szomszor et al., 2010, Carlson et al., 2010), and measure health conditions across time and geographic location.

Significant correlations have been found between social media data such as Twitter and Facebook and clinically validated data sets (Paul et al., 2011; Paul et al., 2014; Achrekar et al., 2013; Broniatowski et al., 2013; Culotta, 2010; Gesualdo et al., 2013; Chew et al., 2010). In addition to the comparable performance of social media based surveillance systems to existing surveillance programs, social media surveillance has the added advantage of speed and reduced costs (Bernado et al.. 2013).

Several studies found significant correlations between social media data and clinically validated surveillance systems such as the Food and Drug Administration (FDA) and the Centers for Disease Control (CDC) for diseases such as HIV (Young et al., 2014), SAR (Yom-tov et al., 2015), adverse drug reaction (Friedfeld et al., 2008; Yang et al., 2014), foodborne disease (Newkirk et al., 2012), suicide (Jashinsky et al., 2013), allergies (Lee et al., 2015), obesity (Gittelman et al., 2010; Chunara et al., 2012), cholera (Chunara et al., 2012), depression (DeChoudhury et al., 2013) and conjunctivitis (Deiner et al., 2016).

The most frequent illnesses for which there were significant correlations between social media data and clinically valid datasets were influenza related illnesses (ILI). These ranged from a significant correlation between blog posts containing influenza keywords and CDC ILI surveillance data (Corley et al., 2010), search queries and ILI illness rates (Woo et al., 2016), Wikipedia articles and CDC ILI activity levels (McIver et al., 2014), influenza related tweets and clinical surveillance gold standards (Guy et al., 2012, Yom-tov et al., 2014, Kim et al., 2013), including Japan's Infection Disease Surveillance Center (Aramaki, 2011), laboratory confirmed influenza cases (Anoshe et al., 2014), and China's Baidu Attention Index (Guy et al., 2013).

However, it is worthwhile to note an outlier — a study in Turkey did not find a strong correlation between national health records of ILI and online data such as tweets, news articles and Google search trends (Blige et al., 2012). This raises a potential research question on how the correlation between social media data and health outcomes may differ regionally. Not only did social media enable real-time assessment of ILI, it also enabled forecasting of ILI. In comparison to a baseline using historical data, Paul et al., (2014) found that Twitter reduced forecasting error by between 17 and 30%.
Tracking public health attitudes and behaviours: The literature reviewed also contained examples of using social media to gauge public health attitudes. Some of the reviewed articles speculated on the use of digital data to track patient experiences and quality of care (Rozenblum et al., 2013; Greaves et al., 2013) as well as to track disease-relevant behaviours and sentiments (Salathe et al., 2013). There were also studies to this effect. Larson et al. (2013) were able to disaggregate vaccine related tweets to provide real-time analysis of vaccine concerns, which can strengthen immunization efforts.

Similarly, Parker et al. (2013) used a combination of health-related tweets and Wikipedia articles as a framework for identifying public health concerns in the population. Chew et al. (2010) saw a significant increase in tweets referring to “H1N1” over time, indicating growing public adoption of World Health Organization recommended terminology for swine flu. A correlation was also found between positive sentiments toward tobacco and individuals’ social image, personal experience and the popularity of tobacco products such as hookah and electronic cigarettes (Myslin et al., 2013), thus suggesting social media as an entry point for tobacco cessation. A computed social media alcohol index correlated with ground truth from the United Kingdom’s Health and Social Care Information Centre weekly alcohol consumption patterns (Kershaw et al., 2014).

There was also a correlation between flu case data from the World Health Organization and tweets that were self-protective in response to epidemics, such as those regarding increased sanitation or avoiding gatherings (Collier et al., 2011). Finally, Hanson et al. (2013) found a correlation between twitter users mention of prescription drug abuse and that of their network. These examples provide preliminary reason to expect that social media data such as tweets can be used to track behaviours such as abuse of alcohol, abuse of prescription drugs, self-protective behaviour in response to influenza, attitudes to tobacco use, patients’ perceptions of quality of care and patient’s level of awareness of epidemics. These data strengthen the need to explore social media as an avenue for community and behavioural health monitoring and potentially interventions.
Conclusion

This study provides a review of literature on social media as a surveillance tool for public health. Based on the literature, social media can be an accurate, timely, granular and cost-effective means for tracking population health sentiments, behaviours, outcomes and emergencies. By democratizing health information sharing, social media allows for collaboration between the public and public health authorities. In light of the reviewed literature, this chapter will discuss future directions in the use of social media for public health surveillance, identifying opportunities for advancing knowledge in the field as well as challenges that need to be overcome to do so.

Strengthening spatiotemporal forecasting models: A knowledge gap in the field of social media surveillance is in the use of social media not only to assess current health issues but also to forecast health conditions (Paul et al., 2016; Culotta, 2010). Further research and improvement of forecasting capabilities as well as data collation to serve is needed to serve this purpose (Chakraborty et al., 2014). Already, models for social media forecasting are being tested and demonstrating effectiveness (Zhao et al., 2015; Culotta, 2010).

Building infrastructural capacity for big data: Big data, such as those harvested from social web applications (Blise et al., 2012) allow for an integrated perspective on global health (Brownstein et al., 2008), including information for social media sites, medical records, purchases, and mobile phones (Bill et al., 2013). This allows for the exchange, analysis and visualization of a diverse amount of information thus supporting situational awareness of health issues (Brownstein et al., 2009). In line with this, there is need to further develop and demonstrate systematic procedures for assessing health information (Stefanis et al., 2013), handling large amounts of data and preventing an overload of information due to noise (Velasco et al., 2014). This will require running more effectiveness studies on social media surveillance systems as well as a structured evaluation of integration systems (Velasco et al., 2014).

The potential of social media as an intervention site: Future studies and interventions are needed to explore the role of social media as an intervention site, given its usefulness in assessing public attitudes and sentiments regarding health issues. This will include shifting from monitoring behaviours to inducing changes in behaviour for health related practices that have been shown to be modulated by social influence (Hill et al., 2013). The term socially shaped diseases or social diseases has been used to refer to diseases for which the burden of disease is spread through social networks (Coiera et al., 2013). These include behaviours such as smoking, poor medication compliance, health seeking for depression, exercise, alcohol sales volume, and drunk driving (Hill et al., 2013; Culotta, 2013). The social nature of such diseases creates an opportunity to use social media to provide network therapy, in line with social network theory (Coiera et al., 2013).

Strengthening data mining and data validation: While traditional surveillance means face the challenges of high costs, imprecise resolution and the possibility of bias (Chunara et al., 2013) social media surveillance is not without its own challenges, especially regarding validity. Most times, the number of geotagged tweets is few relative to the volume of tweets (Kumar et al., 2015). In addition, only a subset of the population is on these platforms. For example, Twitter data tends to be in the age range of 5-24, and 25-49 (Achrekar et al., 2012). There is therefore need for more granularity in social media based surveillance systems.
There is also a need for scalable data infrastructure with capabilities to refine data using more effective machine learning systems (Brownstein et al., 2008; Bernado et al., 2013; Chakraborty et al., 2014; Collier et al., 2011; Hwang et al., 2013; Kostkova et al., 2013; Schmidt, 2012; Salathe, 2016). These systems will need to be able to handle uncertainty, reduce noise, handle unstructured biased data, and be flexible to changes in vocabulary (Brownstein et al., 2008; Bernado et al., 2013; Chakraborty et al., 2014; Collier et al., 2011; Hwang et al., 2013; Kostkova et al., 2013; Schmidt, 2012; Salathe, 2016).

Big data analytics are needed to handle the large volumes of data obtainable from social media (Kumar et al., 2015). This will also include developing disease specific lexicons and more studies on language processing while measuring variations in the effectiveness of machine learning systems across diseases. Nagel et al. (2013) found a stronger correlation coefficient in analyzing social media surveillance for flu when compared to social media surveillance for pertussis. In addition they found a difference in the correlation coefficient between flu tweets and influenza tweets and whooping cough tweets, demonstrating a need for more understanding of the roles that keywords play in the effectiveness of surveillance systems (Nagel et al., 2013).

Most of the literature on the effectiveness of social media for public health surveillance is still observational rather than systematic and predictive (Capurro et al., 2014). There is therefore a need to evaluate the effectiveness of social media for public health monitoring and communication using robust and clearly stated procedures and hypothesis testing so that social media can be integrated into health data systems (Chuanara et al., 2012; Bansal et al., 2016; Brownstein et al., 2013; Kass-Hout et al., 2011; Paul et al., 2016; Padmanabhan et al., 2013; Yang et al., 2013). This will also include investigating the validity of findings across geographical and cultural contexts. As seen in this review, many of the studies were in the United States and only few of them focused on other countries such as Turkey, China, South Korea, United Kingdom, Haiti, Mexico, Australia, Latin America.

**Ethical issues:** The use of social media as a surveillance tool also raises ethical concerns. One of these concerns is related to privacy, the right to use the public’s information when consent is not expressly given, and the expectations of the public on how their information is being used in public health research (McKee, 2013; Moorhead et al., 2013; Fung et al., 2015). Additionally, given the democratization of information sharing using social media, it is possible for inaccurate information or information based on vested interests such as that of industry to be shared with people seeking health information on the internet, thus creating an accountability and authenticity challenge (Greene et al., 2011, Coiera et al., 2013).

**Integrating social media data with traditional public health surveillance infrastructure:** Future directions in this field point to social media as a complement to traditional systems, strengthening their capabilities to gather timely emergency, public sentiment and epidemic intelligence (Yang et al., 2013, Rozenblum et al., 2013, Milinovich et al., 2014, Denecke et al., 2012, Achrekar et al., 2013, Bernado et al., 2013). 75% of the authors in a review of studies in the field called for social media to support existing systems (Bernado et al., 2013).
While traditional case-based surveillance systems such as electronic health records have high veracity, they lack granularity and timeliness, while patient generated data such as social media have high velocity variety without the scientific level of validity and reliability (Salathe, 2016, Kostkova et al., 2013). These systems thus have complementary strengths and researchers like Lee et al. (2013) have called for an “all data revolution” rather than a “big data revolution” with online sources filling gaps in epidemiological data by providing large sample size and more granular spatiotemporal insights (Dechoudhury et al., 2013; Bansal et al., 2016, Chunara et al., 2013; Brownstein et al., 2009, Brownstein et al., 2008; Gittelman et al., 2010). In light of this, more research, funding and partnerships are needed on integrating social media with traditional surveillance systems and building health worker capacity for such levels of analysis.

**Strengthening lexicon development and models for text mining and classification:** Finally, there is need to strengthen best practices in data mining for public health surveillance. This includes further research on natural language processing and parts of speech tagging to reduce noise, increase sensitivity and extract useful information (Kumar et al., 2015; Aramaki, 2011, Cassa et al., 2013; Dredze, 2012, Doan et al., 2012, Anoshe et al., 2014). Other best practices for which more structured best practices are needed include geocoding to enable tracking of point of origin and propagation rate of disease symptoms (Brownstein et al., 2009; Nagel et al., 2014; Lampos et al.; 2010; Kumar et al., 2015).

Machine learning models for data extraction show promise in that they yield significant correlations with ground truth (Lampos et al., 2010). General purpose models such as the Ailment Topic Aspect Model proved effective in identifying health topic models, with statistical significance to ground truth (Paul et al., 2014). More research is needed on noise removal procedures such as in the use of algorithms for automatic classification to reduce error and false positives (Gesualdo et al., 2013, Sarker et al., 2015, Culotta, 2013, Culotta, 2010, Cassa et al., 2013, Broniatowski et al., 2013; Cassa et al., 2013).

Investments in strengthening text discrimination can improve the validity and reliability of social media surveillance systems. This includes through noise filtering, content analysis and categorization, text classification and the use of graph based mining (Blige et al., 2012, Scanfeld et al., 2010; Lamb et al., 2013, Marsen-Haug et al., 2007, Freifeld et al., 2008, Stewart et al., 2011, Aramaki, 2011, collier et al., 2011, Anoshe et al., 2014; Corley et al., 2010). Finally, continual advances are needed in lexicon development and testing the comparative effectiveness of dictionaries for content analysis (Chakraborty et al., 2014; Cassa et al., 2013).

In conclusion, social media as used by surveillance platforms like Sickweather provides the opportunity to track health conditions in a timely and cost-effective manner, especially for outbreaks for which there is no cure. Social media can allow for the spread of information, for collaboration and for effective detection and monitoring of outcomes. Future research is needed to develop and confirm structured procedures for mining social media data, to build health systems capacity for big data and to test the use of social media in forecasting health trends as well as a potential intervention site for public health.
References


REFERENCES (c.)


REFERENCES (e.)


About The Author

Ebele is a Doctor of Public Health with a specialization in Community and Behavioural Health. She has a Masters in Global Health and Public Policy and an Undergraduate degree in Biomedical Science with minors in Biology and Psychology.

In addition to her work with Sickweather, she has worked on innovative health projects globally for WHO, UNICEF and also serves as President of the Engage Africa Foundation.

Ebele Mogo, DrPH
Public Health Policy Consultant