INFODEMIOLOGY TO IMPROVE PUBLIC HEALTH SITUATIONAL AWARENESS:
AN INVESTIGATION OF 2010 PERTUSSIS OUTBREAKS IN CALIFORNIA, MICHIGAN AND OHIO

Jennifer Olsen

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Public Health in the UNC Gillings School of Global Public Health.

Chapel Hill
2013

Approved by:

Courtney D. Corley, Ph.D.

Kristen Hassmiller Lich, Ph.D.

Thomas Ricketts, Ph.D.

Richard Gary Rozier, D.D.S.

William Craig Vanderwagen, M.D.

ABSTRACT

JENNIFER OLSEN: Infodemiology to Improve Public Health Situational Awareness: An Investigation of 2010 Pertussis Outbreaks in California, Michigan, and Ohio
(Under the direction of Thomas C. Ricketts)

As a disease emerges, one of the greatest challenges for public health practitioners is to differentiate between a normal event and a serious outbreak. Typically, information from official sources and surveillance systems had been the only resource. More recently, the field of infodemiology has emerged with a focus on the distribution and determinants of health information on the Internet. This research compared official reports of whooping cough with infodemiology sources, specifically news articles, search engine patterns, and Twitter, to assess the timeliness, accuracy, and correlation of these content sources. Within California, Michigan and Ohio, Internet search patterns identified the outbreak of pertussis in 2010 four to eleven weeks in advance of official sources, and there was strong correlation between the epidemic curve and search pattern in Michigan and Ohio. Twitter also provided an indicator in advance of official sources in all three states, but only with a single Tweet. Using all three sources to identify indicators was better than any single source used independently.

While understanding the data utility is important, it is equally critical to understand the attitudes and perceptions amongst public health leaders regarding the use of infodemiology data to improve situational awareness. A survey of such leaders showed that infodemiology content had the most value in the first stage of situational awareness for identifying early indications of disease outbreaks. News media and Internet search were

moderate to highly valuable for 70% of respondents, while social media was moderately to highly valuable to 60% of respondents. For both strengthening the comprehension of an outbreak and informing future predictions, beliefs were split regarding the level of potential value (if any) that exists. This led to a framework on how to include infodemiology content in public health situational awareness strategies going forward, so limited resources are used as effectively as possible.

ACKNOWLEDGEMENTS

My personal support team has been an impetus for the beginning and sustaining of this project and to continue bringing this work to a close. I am especially thankful to my very patient husband, Ken, who has been my cheerleader, support staff, and sounding board through each step of this process. I also must recognize my parents and sister, who have sat through years of listening to me prattle on about diseases, disasters, and my dissertation. I know they are in my corner constantly, making sure I achieve any goal that I set myself. And then there are my friends, who have gotten used to me answering "No, I'm ‘dissertating’," but they have not given up on me.

My classmates in this program have become more than a random cohort; they are my friends and my teachers. Our cohort (The Sixers) is fortunate to have become a strong, cohesive unit, rooting each other to the finish line. I would have never met this accomplished group had it not been for Amy Kircher, DrPH and Jean O'Connor, DrPH who tirelessly 'suggested' I should look into the program at UNC and that it would be worthwhile. Their 'gentle nudging' and help along the way has been invaluable to my success.

Another group that has been a strong supporter of this dissertation in numerous ways are my colleagues within the Office of Preparedness and Emergency Operations (OPEO) in the Office of the Assistant Secretary of Preparedness and Response (ASPR) at the U.S. Department of Health and Human Services (HHS). Many times I have had to finish work late, needed someone to cover for me, and just be a little less available than usual to get through this program. I truly appreciate their unwavering support over the past three years.

I must also thank the International Society of Disease Surveillance (ISDS), who was kind enough to distribute my survey to their membership. I recognize there are many surveys that any professional society is asked to share, and their belief in the value of my work is truly appreciated. I also owe a big thank you to my two data providers: HealthMap (namely John Brownstein and Sumiko Mekaru) and Gnip (Mark Hoy and Stuart Shulman).

Last, but not least, I owe all my gratitude to my dissertation chair, Dr. Tom Ricketts, who listened to my early concept as it unfolded and guided me through this dissertation journey. Some of my approaches have been unusual (reading Twitter), some of my reasons for being behind deadlines have been outlandish (Superstorm Sandy), but through all of this, Dr. Ricketts has supported and enabled this work to continue. I am also extremely grateful to the rest of my committee: Dr. Corley, Dr. Lich, Dr. Rozier, and Dr. Vanderwagen for their thoughtful guidance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1. BACKGROUND

The emerging field of infodemiology offers promising tools that can help public health leaders streamline the process of monitoring, processing, and utilizing unofficial sources to aid in their decision-making. Infodemiology is "the science of distribution and determinants of information in an electronic medium, specifically the Internet, with the ultimate aim to inform public health and public policy" (1). It allows practitioners to track what community members who are not public health experts communicate using the Internet in order to measure the public's opinions, attention, behaviors, knowledge, and attitudes (1). Communication from unofficial sources, such as the Internet, may well-inform decision makers and enable others to respond proactively.

Infodemiology may be critically valuable during a disease outbreak, during which accurate information can improve decisions that can then save lives because it enables more reliable event detection and timely response. Event detection is the term used to describe the differentiation between baseline occurrence of disease and more severe outbreaks that would require a rapid response (2). Early detection of disease reduces negative impacts because this detection allows for the implementation of timely interventions. Effective event detection also helps in decision-making regarding whether further investigation is required.

Because event detection is so important, time is a critical factor before and during disease outbreaks. Each minute in the early stages of disease detection matters because many rapid, critical decisions are required which are based on rapid information sharing. In the early stages of a disease outbreak, there is an overwhelming amount of information and misinformation that begins to emerge about the event, its causes, and the impacts of the

disease occurrence. Infodemiology, to be effective, has differentiate between the bad and the good that appears on the Internet, as well as fuse with information coming via “normal” channels. Incorpoating infodemiology data into public health practice may improve detection and intervention.

At an early stage of an outbreak, the significant challenge of managing information effectively during a disease outbreak occurs. Information is imperative to detecting and intervening in a disease outbreak, but the enormous amount of information that public health leaders must sift through in this limited amount of time can become overwhelming and obstructive. These overwhelming amounts of information are produced by a variety of sources (governments, news organizations, citizens) and this information is often conflicting, repetitive and/or erroneous. Because of these information challenges, public health leaders (who serve as information consumers) need to understand the usefulness of such content when there is limited time to read, comprehend, and process information for decision-making. Therefore, this is where infodemiology may prove useful.

In the disease outbreak environment, it is imperative to understand which sources of information add value and should be used for decision making in this limited timeframe. It is also important to understand whether or not unofficial sources can provide insight in advance of official sources and if they can be used with confidence. However, monitoring of information sources is a time consuming and resource intensive activity, so any ways to simplify the process by focusing on sources with proven utility is beneficial.

Research Question: ***Can infodemiology improve public health situational awareness?***

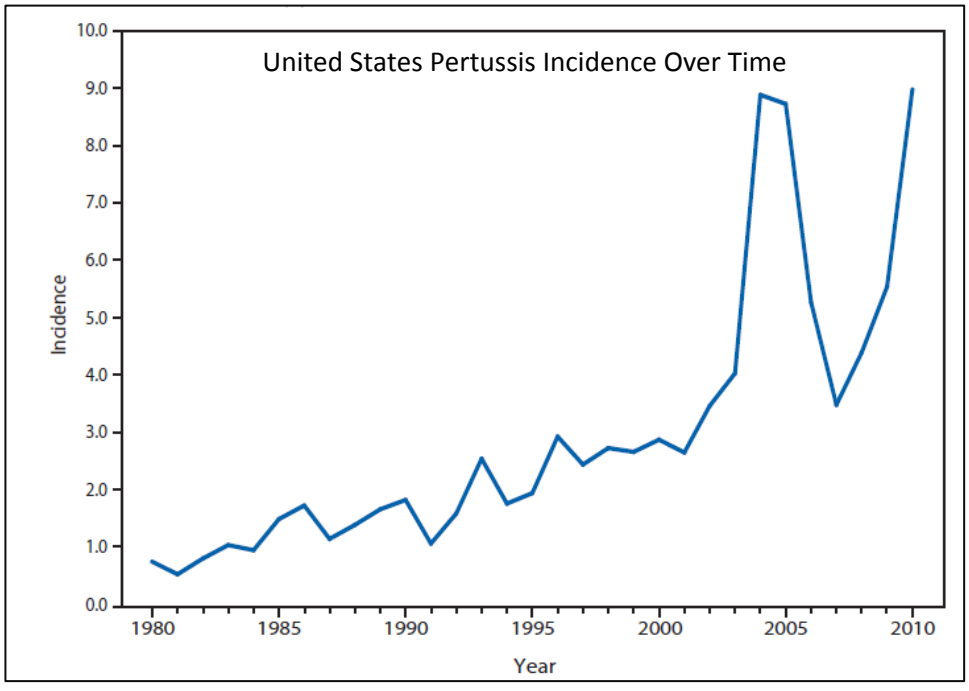
Null Hypothesis: *Utilization of infodemiology data does not improve situational awareness during an outbreak of pertussis.*

Pertussis

Pertussis is an acute bacterial disease, often contracted by children, caused by *bordetella pertussis,* which infects the respiratory tract and often lasts one to two months (3). The disease often starts with a cough that worsens, and the cough may lead to a characteristic whooping sound upon inhalation (3). In the United States, 80% of deaths from pertussis are amongst children under one year of age, with pneumonia being the most common cause of death (3). The incubation period is seven to twenty days, and the communicability period is three weeks (3). Transmission occurs via direct contact with mucous discharge via airborne droplets (3).

The total number of pertussis cases nationwide in 2010 was 27,550 as compared to 16,858 in 2009 (4). Nationwide, pertussis has a cyclic peak every three to five years; however, the incidence in 2010 surpassed the peaks experienced in 2004 and 2005 (Figure 1.1) making it a notable year for the disease with an incidence of 8.97 cases in every 100,000 people (4).

Figure 1.1. *United States pertussis incidence (per 100,000 population) by year (4)*



In 2010, large localized outbreaks of pertussis occurred in California, Michigan, and Ohio (5). When compared to the previous year, these three states had a significant increase in annual case count.

Table 1.1. *Difference between case counts in California, Michigan, and Ohio (4)*

State	2009 Total Cases	2010 Total Cases	Delta
California	896	3,080	+2,184 (+244%)
Michigan	900	1,500	+600 (+67%)
Ohio	1,096	1,806	+710 (+65%)

This significant increase (seen in detail in Appendix A) makes pertussis a potential candidate disease to study for this research. Additionally, pertussis is a nationally notifiable disease with associated lab confirmations, so there is higher confidence in National Notifiable Diseases Surveillance System (NNDSS) being a reliable "gold standard" data source. Pertussis does not have a distinct seasonality (6), which the researcher confirmed by examining the past five years of pertussis in the each of the three states (see Appendix A). Lastly, neither strong cultural taboos nor negative implications are associated with pertussis, either of which would cause reporting delays (7).

CHAPTER 2. RATIONALE

Traditional Surveillance

Traditional surveillance is critical for the protection of populations from disease, but surveillance systems are often slow. Most surveillance systems detect signals immediately before or near the time when signals appear in gold standard data (2); this timing is not sufficient to provide an early indicator of emerging outbreaks because of the delay in appearance of gold standard data. For diseases that cause severe mortality, existing approaches to mortality surveillance do not result in disease detection in a timely way, supporting the need for new data sources to track public health impact (8). Syndromic surveillance systems, which focus on using data from chief complaints rather than diagnosis, need additional data streams to increase signal detection sensitivity without decreasing the specificity required (8). Clinic-based syndromic surveillance and microbiological testing for verification and diagnosis are also critical. Limitations of traditional surveillance systems include, but are not limited to, reporting delays, inconsistent population coverage, and poor sensitivity (9, 10).

However, traditional surveillance is necessary for estimating morbidity, mortality, and shifts in disease incidence within demographic or other groupings. In the United States, compliance with notifiable disease reporting rules varies from 9% (Invasive *Streptococcus pneumoniae* in Hawaii*)* to 99% of cases (11) where notifiable diseases are any conditions where "regular, frequent and timely information regarding individual cases is considered necessary for prevention and control of the disease" (12). For instance, the National

Notifiable Disease Surveillance System (NNDSS) run by CDC was evaluated and the median national reporting delay was forty days for pertussis (13). However, this nationwide delay is not necessarily representative of all states, since each state has a sequence of actions they take prior to reporting to a nationwide system (to include various policies and protocols) (13). In addition to state-variable reporting timelines, some states only report lab-confirmed cases, which further delays reporting (13). Another limitation of traditional approaches is that both sentinel surveillance systems and laboratory systems will likely over-report incidence in groups who are more vulnerable to the disease, therefore, making alternative surveillance methods more critical for accurate detection (14).

Due to this inherent variability between states and across diseases, the ability to depend on weekly reports from a system like NNDSS is limited, especially for multi-state outbreaks (13). Fortunately, publicly derived unofficial reporting of outbreaks can be faster than official channels and, at the same time, it can be reliable and responsive to the needs of local public health workers (15). Novel Internet based collaborative systems can have an important role in gathering information quickly to improve coverage, accessibility, scalability, timeliness, and transparency of traditional surveillance systems (16). Infodemiology data (often derived from Internet blogs, websites, query information and navigation data) can be collected and analyzed in near real-time, giving public health leaders the opportunity to put a finger on the pulse of public opinion, behavior and knowledge (1).

The Federation of American Scientists and the World Health Organization (WHO) co-sponsored an event in which a group of conference attendees created ProMED-Mail (Program for Monitoring Emerging Diseases), which analysts have used as both a nontraditional information source and a personal network to share information which provides an early warning system for outbreaks to protect global public health (15) as early as 1993. Volunteer experts across the globe moderate ProMED-Mail, a freely available non-government system, in which local observers, news reports, and other content streams

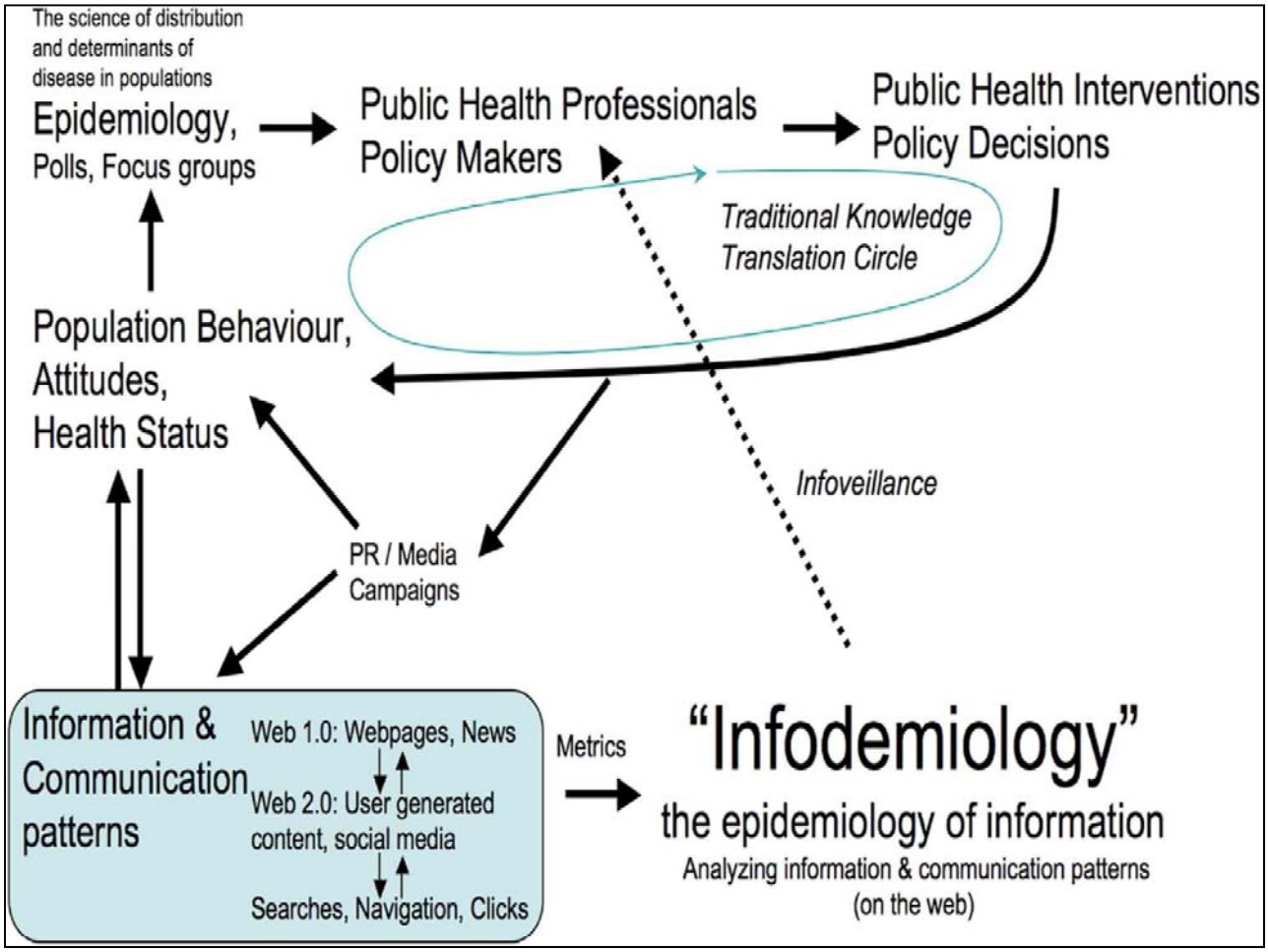
share information. The WHO has credited ProMED-Mail as the source of reports on various diseases, including the outbreak of an undiagnosed disease in Pakistan in spring 2000, which was later confirmed as pulmonary tuberculosis (15). Severe Acute Respiratory Syndrome (SARS) first became known to the Western world through a ProMED-Mail report. This report is credited with breaking the story and providing critical information that helped to rapidly identify the cause of the disease (15). The United States government endorsed the open source work of ProMED-Mail in 2001. D. A. Henderson, at the time, served as Health and Human Services Secretary. Tommy Thompson's principal science advisor for public health preparedness, congratulated the publication on being one of his office's primary sources throughout the September 11th anthrax attacks (15).

More recently, research has found that social media sources and news media sources may provide indicators of disease outbreaks prior to traditional reporting sources (i.e. surveillance systems) (17, 17). In fact, the WHO uses informal information sources for about 65% of their outbreak investigations and relies on informal sources for daily surveillance activities (18). Surveillance systems that utilize informal Internet-based information have been shown to reduce the time to recognition of an outbreak and facilitate responses to disease outbreaks (19). Additionally, persons' use of search tools such as Google® to seek information about a suspected disease or symptoms may be indicative of an emergence of a disease, and web searching may also provide early indicators of a disease (20). Furthermore, an analysis of this search tool data has the potential to capture information about people who do not seek formal medical care (21).

Apparently, the dynamic nature of the continuously updated "social web" makes it a fertile environment for intelligence gathering in a variety of disciplines, (22) enabling public health leaders to tap into the "wisdom of crowds" (via crowdsourcing). Because of this source, the public plays a potentially larger role in all stages of knowledge translation, which

includes information generation, filtering, and amplification (23). In a public health context, infodemiology can empower the lay public as a source of information (1). This information can then become the content that influences other people. Additionally, because of the way that this content is collected and distributed, it can have an impact and value at the population level (1). Thus, all of these elements contribute to a useful, potentially critical framework to utilize infodemiology (1) when preparing for and responding to disease outbreaks [see Figure 2.1].

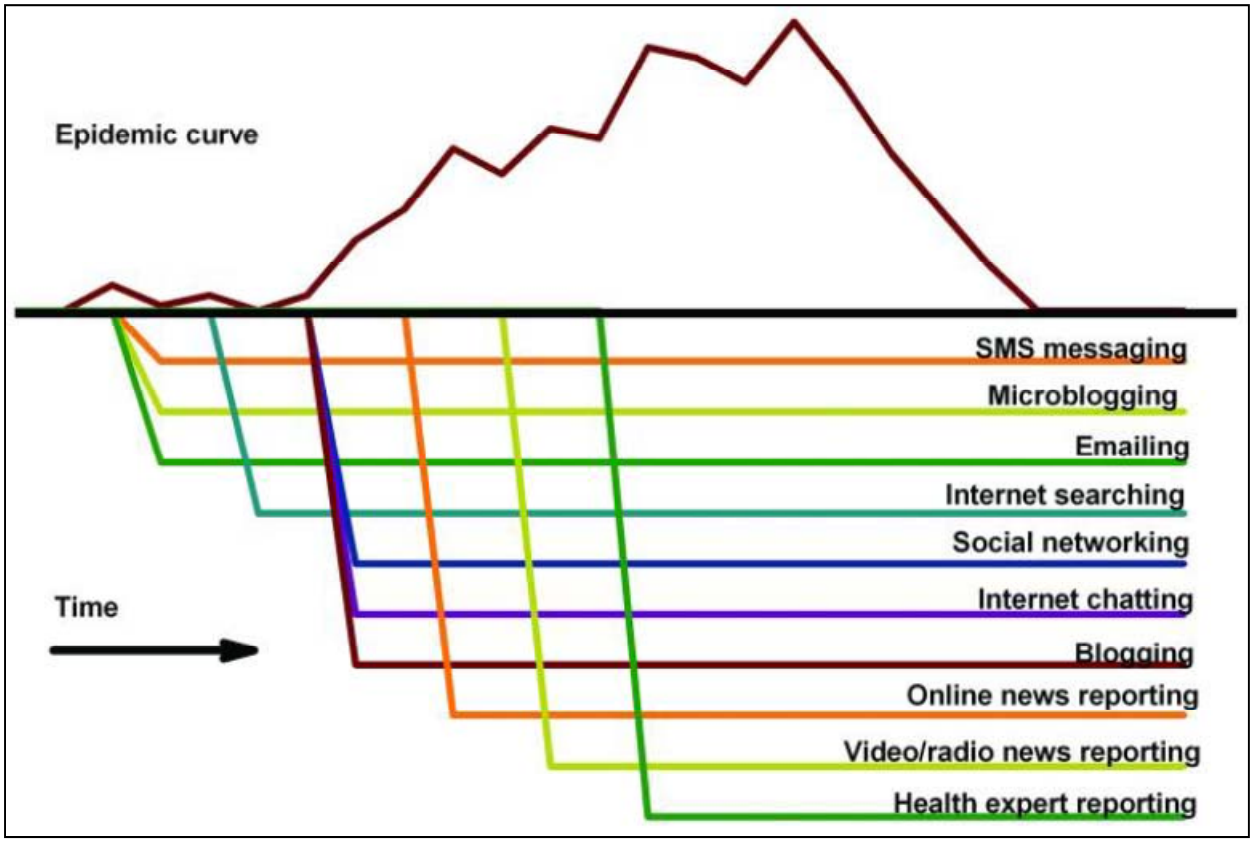
Figure 2.1. *The role of infodemiology in public health (1)*



Since it has been shown that a single infodemiology data source can inform us of an emerging epidemic prior to official event confirmation [the timing of the appearance of

mentions or indicators of a disease are illustrated in Figure 2.2], then it would seem feasible to determine if these same sources, when combined, can be used to provide worthwhile or even early information during an outbreak. Most of the surveillance systems that were evaluated only used a single data source (2), making this multi-source investigation novel and important.

Figure 2.2. *Potential information timeline for disease surveillance (104)*



News Articles as a Source for Infodemiology

Newspaper coverage both influences and is influenced by epidemiological reports (7) because it distributes information about disease while also offering new insight into the disease and its effects on populations. In this way, newspaper coverage is a great example of how unofficial information spreads in a disease outbreak. Studying the news as an

indicator of public health impacts provides insight into disease impacts on individuals and society in a way that is very different from official reports. Unlike these reports, news may validate the spread of a disease or condition, especially if it is something new for the community (24). Additionally, more than 64% of Americans have used online news to get information about current events (25), making web-based news a critical information sharing tool. Due to the pervasiveness of web-based news, newspaper publishers and owners are shifting their focus from providing content in paper form to an electronic form to their readers.

News media reports encompass both direct indicators of disease (such as the number of people sick) as well as indirect indicators (such as the lack of available remedies) of societal response produced in real-time; this adds value beyond traditional approaches (17). Newspapers also have significant flexibility regarding what aspects of an event they can cover and, therefore, they are a good indicator of the general public's interests and concerns (7, 26). Significant public concerns or fear may impact the news content, (7), can be indicative of a new health issue (24), and can identify unusual outbreaks (10). The World Health Organization (WHO) reported that between July 1998 and August 2001, 56% of the 578 verified outbreaks were first identified by the Global Public Health Intelligence Network (GPHIN), which primarily uses newspaper content (18).

One of the preeminent tools for monitoring, organizing, harvesting, and visualizing news articles related to emerging infectious diseases is HealthMap (27). This system automatically scrapes the web, with a focus on various Internet news sites around the globe, and then has expert staff curate and annotate the stories so they are freely available to the public. Text processing algorithms run constantly to produce alerts classified by language and disease, as well as maps with customizable views (28). Since 2006, Brownstein and Freifeld have been improving the utility of news articles for public health disease

surveillance and public health preparedness by turning information into useful knowledge for the public, and public health officials (9, 28, 29).

Internet Search Patterns as a Source for Infodemiology

The pattern of how and when people search the Internet has the potential to provide clues or early indicators about future concerns and expectations amongst the public (30), and has shown initial promise as a source of infodemiology content. Search query data is more timely than traditional surveillance; often taking a process that has been years in duration and moving it to real-time (31). For several reasons, it is also more efficient: it can be automatically collected in a centralized fashion and shared with officials; it does not require reporting infrastructure in each region (31); and it can be more cost effective (32). This collective phenomenon of Internet search behavior is a meaningful, robust reflection of human behavior and disease patterns across large populations (33) and the information can serve as a surrogate for traditional measures of disease burden (34). Additionally, Internet search patterns are just one potential measure of "social proof" which can determine that if many people do or think similar things (i.e. search for the same things on the Internet) then, there must be good reason (35).

Of Internet searches done in December 2011 within the United States, over 65% of those were done using Google (36). In the month of December 2011 alone, more than eighteen billion web searches were conducted in the United States, with over twelve billion of those searches done on Google (36). Over time, Internet use will expand to be a more integrated part of individual lives and the Internet will be used by more representative parts of the population as a whole (14). Because of this, Google can utilize the collective intelligence of millions of web user logs for disease monitoring, as evidenced by influenza efforts (37) and the eventual development of *Google Flu Trends (http://www.google.org/flutrends/us/#US).* Additionally, web queries enable access to

individuals who are ill but who are not (or not yet) seeking care, especially during the early stages of an illness (21); therefore, providing an accurate, low resource mechanism for surveillance (14).

In 2010, 59% of American adults sought health information online (38), and 37% accessed user generated health information online (39).This includes web 2.0 content where users supply the content that other users demand (32). To start their searches, 66% of health seekers begin with the use of a search engine (40), making search information a generalizable marker. There is some socio-demographic variation amongst those who utilize the Internet for health information as compared to those who do not: women, people with a college education, those with higher incomes, and younger people are the most likely to utilize the Internet to access health information (38). Searches for health information occur at about the same level of popularity as paying bills online, reading blogs, or looking up phone numbers and addresses, making it extremely common (40).

The online conversation about health-related topics is being driven by the availability of social web (or web 2.0) tools and the motivation of people to connect with one another, especially amongst those living with chronic conditions (38). The proportion of people who seek health information on the Internet is the best measure for health-oriented search behaviors, especially when compared to indirect measures such as proportion of households with Internet use (41). During 2009 H1N1, respondents cited the internet as their most frequently used source of information about the pandemic (42). Internet search patterns are a more favorable measure for high prevalence diseases where early detection would enable prevention of spread and for diseases where there is fluctuation due to seasonal change or an occasional surge in case counts to enable matching of actual cases to search upswings (21).

Social Media as a Source for Infodemiology

Although individual messages on social media sites contain little informational value, aggregation of millions of such messages can generate valuable insights (43). Tweets can enable real-time content analysis and knowledge collection, allowing health authorities to respond to public concerns (23) much faster than ever before. Twitter content exists within an analytical “sweet spot” in that they are long enough to provide depth and meaning, yet concise enough to facilitate rapid analysis and classification (1), which makes this another potentially valuable source of infodemiology content. Twitter has provided rapid, cheap, reliable content for assessing events from earthquakes to seasonal allergy patterns (44) to the Fort Hood shooting (45). Gupta found that 30% of Tweets posted about an event contain situational awareness content, with only 14% of the Tweets containing spam (46), while Vieweg determined that 8-24% of Tweets contained tactical, actionable information in recent events (47). This type of information may be useful for directing limited resources to reduce impacts (48), to include morbidity and mortality. Twitter enables people to rely on a crowd rather than an individual, which increases their chances of finding information they did not know existed (35). This information utility is especially true in the United States where "users in the US give Twitter a more informative purpose" (49).

Additionally, distributed networks of concerned citizens (forming notional crowds) to share situational awareness use social network tools (50). These groups share information both horizontally (peer-to-peer) and vertically (to organizations involved in the event) (50). Groups are intelligent, and often smarter than the smartest individual within the group (35). The simplest way to get reliably good answers to a question is to ask a large and diverse group every time, because even if most people in the group are not well informed the collective intelligence is excellent (35). Horizontally shared information is more timely, complete, and of a higher quality (better sensitivity and specificity) (50). This information

quality is especially true in communities where local populations know the baseline events in their community and can rapidly identify anomalies and network to other neighbors to rapidly get “ground truth” (50); through decentralization they are able to draw on local insights (35). Social media has become a valuable resource since it can provide content that is not otherwise available through traditional information management. It is especially valuable because the user community is self-policing to reduce misinformation (51) (see Table 2.1).

Table 2.1. *Traditional and nontraditional media characteristics, adapted from Keim and Noji (51)*

	Traditional Media	**Social Media**
Information Flow	Single direction (from media to public)	Multidirectional
Information Control	High	Low
Adaptability	Low	High
Local Relevance	Low	High
Information Accuracy	Variable	Variable
Timeliness	Delayed	Immediate

Twitter is becoming a key part of the way people use the Internet with 16% of online adults in the United States using the tool in June 2011 (25). There is variation by race in Twitter usage amongst Internet users, with black non-Hispanics, having the highest overall usage rates at 25% and the highest typical day usage at 11% (52). There is also variation with the geography of Twitter users as compared to non-users, with urban (15%) and suburban (14%) areas having significantly more users than rural areas (7%) (52). New information sources such as Twitter may be able to provide insight in areas where there is currently a shortage in useful information flow, specifically in areas with a high proportion of black or Hispanic citizens whose use of Twitter is significantly higher than that of white citizens (52). The largest growing group of Twitter users is individuals ages 25-34, with usage rates doubling from 9% in November 2010 to 19% in May 2011 (52). Additionally, because of the increased availability and decreasing cost of cell phones, more than half of all Twitter users access the service on their phones (52). This access means that more infodemiology content will be available as smartphones become more pervasive across all

socio-demographic groups which further points to the increasing need for research and standard analytical approaches.

Twitter users are disproportionately from lower income households and are more ethnically and racially diverse than the general United States population (53), making it an extremely valuable source to understand the health concerns among more vulnerable populations and in areas where the disease burden may be more significant. Additionally, since less than 10% of Twitter accounts are private and unavailable for public review and analysis (54), most accounts and their related content is available for analysis and utilization which reduces the potential for bias between public and private account content.

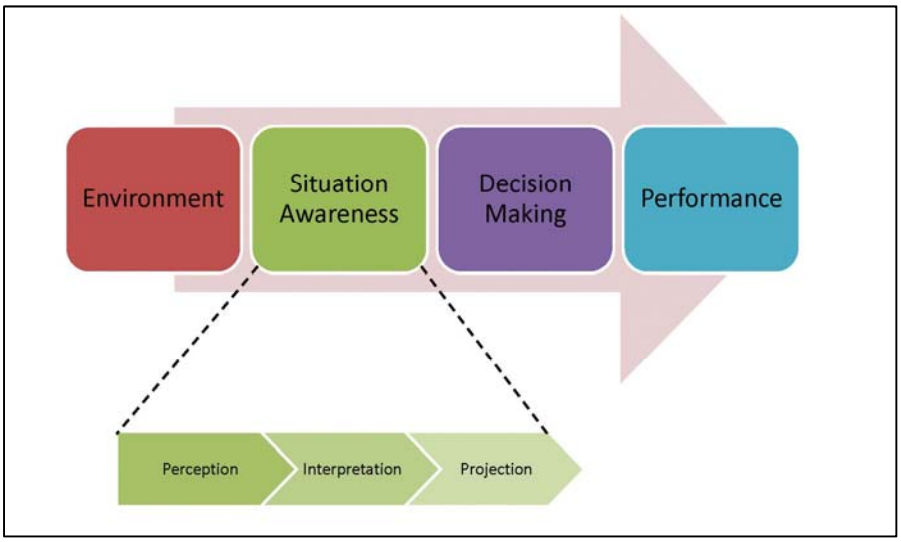
The potential for misinformation on Twitter may be limited as compared to other social networking sites due to the self-policing within the Twitter user community. In the work of Chew, only 4.5% of Tweets included possible misinformation or speculation (23). This limited misinformation possibly adds to the potential utility of Twitter as a source for infodemiology. Twitter is also a very important way for citizens to share information, with 25% of online citizens (and 14% of general citizens) posting information or photos about an emergency or noteworthy event (25), which shows that the public sees Twitter as a way to share eyewitness reports.

CHAPTER 3. CONCEPTUAL FRAMEWORK

A key element in rapidly identifying diseases is situational awareness, and this dissertation examines incorporating novel data sources to potentially improve that awareness. Situational awareness allows public health leaders to know what the current status of disease incidence is in any given area; therefore, it is important in disease identification, outbreak investigation, and prevention. Improved situational awareness can lead to better decision-making and performance (55), which will facilitate all stages of disease outbreak response. To recognize how infodemiology can improve situational awareness, a clearer understanding of situational awareness is needed.

There are three key parts of situational awareness that are illustrated in Figure 3.1: perception of elements in the current situation; comprehension of the current situation; and projection of the future status (55). Relevant and timely information are necessary to inform situational awareness, and good situational awareness is necessary for effective decision-making (50).

Figure 3.1. *Situational awareness feedback loop* (54)



One element of improving situational awareness is effectively evaluating which, if any of these new, nontraditional sources of information have utility. New technologies and information sources are often viewed with skepticism, and many, especially those who work for government organizations, consider information from non-government sources unofficial, and/or unreliable. With this mindset, it seems critical to evaluate the new sources discussed here in order to identify if or when they have utility or value. In addition to evaluating the data sources, it is also critical to examine the perception and the potential inclusion of these data sources into the decision cycle. These new sources, like accepted forms of disease monitoring content, can be categorized as either information that is useful to predicting the future or information that is not useful, potentially erroneous, or even harmful (57). Since the future is unknown, it is impossible at the moment of an outbreak to distinguish which category each source of information falls into so complex decisions often result in the inclusion of information that is not useful (57).

To understand how infodemiology sources can be as or more useful than "official" data collection, we must first understand more about the process whereby we observe, decide, and act on information. Decision-making is a process of pattern storage,

recognition, and prediction rooted in past world experiences and perceptions (56). The brain creates a model of the world, and everything one learns is added to the model, so the brain constantly compares the model to what it sees in reality (56). To make a prediction, the brain will compare past structural knowledge (or patterns) with the most recent details available (i.e., awareness of the current situation) (56). Correct predictions result in situational understanding, while incorrect predictions lead to confusion (56). Information must be analyzed, synthesized, and distributed in near real-time to enable decision-making (50). Successful actions often depend on fast information collected directly from the field ("ground truth") (35).

However, decision-making is often a result of instincts (or gut feelings) that appear quickly in the consciousness. These instinctive feelings usually are due to the underlying feelings that we are not fully aware of at the time. Furthermore, these feelings are often strong enough to act upon and tend to follow "rules of thumb" (57). While these instincts are neither impeccable nor stupid, due to what is called the "beneficial degree of ignorance," they can outperform a considerable amount of knowledge and information (57). Less time and information, at times, can improve decisions, despite two core beliefs in our culture that "more information is always better" and "more choice is always better" (57). People generally believe that the more information the decision makers have, the better off they are. However, extra information can be harmful, confuse the issue, and make people feel more confident than they should. Individuals who constantly overwhelm their brains, and feed it more information than it can handle, may gain only a limited understanding (58).

In an uncertain environment that may be encountered when deciding whether there is a disease outbreak, good intuitions must ignore information. The "quality of intuition lies in the intelligence of the unconscious: the ability to know without thinking which rule to rely on in which situation" (57). There can be as much value captured in the blink of an eye as in

months of rational analysis. Successful decision-making is a result of balancing deliberate and instinctive thinking (59). People who make decisions under pressure, even those considered experts in their respective fields, do not logically and systematically compare all available options (60). The key to good decision making is understanding, and not knowledge (59). Unfortunately, when overwhelmed with information, people drown in knowledge and lack understanding. A conscious brain can only handle five to nine pieces of information at any moment, and problems with more than four variables overwhelm the mind (58). Humans also prefer the illusion of certainty in their information and decision making as opposed to the reality of doubt (35).

Recent studies have shown specific tendencies in the decision making process. As described by Kahneman and Tversky, humans have greater confidence in highly correlated observations, and are often insensitive to the reliability of evidence (61) that those correlations may not be correct. Prediction is the result of selecting an outcome that is most representative of the input data (assuming the input data is error free), and confidence increases with consistency (61). People predict by interpreting the representativeness of observations, utilizing prior or background information or specific evidence about the event, and assessing the accuracy of their prediction (61). Unfortunately, there is little relationship between an expert's confidence in a decision and the accuracy of the decision (35).

The two most frequent cognitive biases in decision making are anchoring: a "shortcut in thinking where a person doesn't consider multiple possibilities but quickly and firmly latches on to a single one" and, availability: the "tendency to judge the likelihood of an event by the ease with which relevant examples come to mind"(62). With anchoring, one of the first potential solutions is latched onto as the answer without continuing assessments. With availability errors, what is most available in the mind strongly colors one's thinking about a new event and makes it seem similar to events one has experienced in the past. This

association causes one to ignore important differences and fail to identify the correct scenario.

Other decision making errors include: representativeness error (thinking is guided by a prototype, so one fails to consider possibilities that contradict the prototype and thus attributes the symptoms to the wrong cause) and confirmation bias ("confirming what you expect to find by selectively accepting or ignoring information") (62). In medicine, physicians are known to fall into the trap of a "zebra retreat," shying away from a rare diagnosis (62). Additionally, for doctors it may be satisfactory to find a cause, but not identifying all potential causes for a patient's ailment may lead to missing a critical event (62). This potential for missing a critical event exists within outbreak detection, requiring diligence by those who receive information. Multiple sources of information about an incident increase the level of confidence about both the event's likelihood of occurrence and various related details, more so than any single report.

These and potentially other, undiscovered processes that relate observations to decisions to action can affect how disease surveillance either succeeds or fails. Situational awareness is a result of the collection and rapid analysis of information sources, and this hypothesizes that infodemiology content is one of the critical situational awareness sources for public health. Following the conceptualized process illustrated in Figure 3.1, improving situational awareness can result in improved decision-making. The following chapters describe how this takes place in actual situations.

CHAPTER 4. LITERATURE REVIEW

To support the work described in this dissertation, a literature review was completed to determine the values of open or social media sources in gathering information on public health topics. The findings indicated that in certain contexts some disease characteristics, their magnitude, and location could be identified prior to official reporting. Forty-one articles were identified in which the authors describe the use of news or social media to gather information about a public health condition.

Methods

Although fields outside of public health may be more advanced in assessing and using traditional media and social media to gather information, the scope of this review will be limited to public health purposes. This literature review will focus on peer-reviewed journal submissions.

Key Words

There are three concepts that needed to be included in the search structure: media (social or news), public health, and input. The key terms in Table 4.1 were used to initiate the search. Truncated versions of the words in Table 4.1, as well as using OR for each term under the category headings were used (i.e. blog OR television AND public health AND report OR assess).

Table 4.1. *Specific search strings used for literature review*

Media (OR terms)		***Public health*** (OR terms)		***Input*** (OR terms)
News*		Public Health		Gather
Media		Health		Collect
Crowd sourcing		Med*		Infodemiology
Twitter		Epidemiology		Infoveillance
Facebook				Surveillance
Blog	AND		AND	Situation* awareness
Television				
Internet				
Web				
YouTube				
Social network*				
Microblog*				

Sources

In identifying possible items for inclusion, databases were tiered for review based on apparent relevance and likelihood of relevant findings. This was done in the beginning in an attempt to review the sources believed to be most plentiful. The following is the tier structure selected for databases to be searched:

Tier 1: *MEDLINE/PubMed, Web of Science (Social Sciences Citation Index), Scopus [via NIH], Communication & Mass Media Complete, Google Scholar*
Tier 2: *Global Health, Embase [via NIH], Academic OneFile*
Tier 3: *IEEE Xplore*

Search Strategies

"Snowballing" was used to identify additional sources that meet the specified inclusion and exclusion criteria. The articles found through this method underwent the same review process as all other articles to identify if they will be part of the final findings. The use of "snowballing" is especially important in a field that is evolving as rapidly as media and technology. The search terms of the past (such as World Wide Web) may seem archaic now, but I wanted to include them since the terminology shift is part of the evolving field. For all articles that deemed appropriate for this review (last criteria of abstraction database), the references cited in those articles were reviewed following the same process as all other articles.

Inclusion Criteria

- Any type of study, assessment or review (including letters to the editor, commentaries, and editorials) that underwent peer-reviewed submission
- Any human study population
- Any location (see note below about language restriction)

Exclusion Criteria

- Articles published in a language other than English
- Articles focused on using media to educate or inform people
- Articles describing mechanisms, methods, or results for using media to recruit participants into studies
- Articles that are duplicates
- Articles, editorials, private blog posts, and findings not within a peer-reviewed journal
- Articles published before 2000 or after April 2011
- Articles about animal or veterinary issues or diseases
- Posters, interviews, and books
- Articles or reports by government agencies, think tank organizations, foundations or academia

Review Process

The goal of this review process was to identify and capture all relevant studies in this subject area for inclusion into the final literature review. The following was my methodology for reviewing articles that had met the inclusion/exclusion criteria and search terms specified above:

- Reviewed all titles of articles that met the criteria and search terms specified. Using subjective assessments, the title was reviewed for topic relevancy. If the title indicated…

    1. Relevancy – the article was included for the next level of review
    2. Irrelevancy in relation to the topic – the article was excluded from further review

3. Nothing definitive on relevancy – the article was included for the next level of review

- Abstracts that met criteria 1 or 3 as specified above were pulled for review. If the abstract met the intent of this review, or if it was unclear, the full article was pulled. If the abstract clearly did not meet the intent of this review, the article was excluded from further review.

- For those articles for which abstracts were identified as relevant or possibly relevant, the full articles were pulled, printed, numbered, and reviewed. All fully reviewed articles were entered into the abstraction database, which included the following fields:

    a) Full citation
    b) Timing of study
    c) Information source analyzed (news, blogs, etc.)
    d) Health outcome of interest
    e) Source data used for comparison
    f) Size of study
    g) Findings
    h) Gaps in knowledge and limitations
    i) Criteria for including and evaluating information sources
    j) Methodological quality
    k) Included for this review (Yes/No)

- A subset of the abstraction database fields were used to generate the tables in this review that summarize findings of articles which were selected for inclusion into this review (and will incorporate findings from step 3 and step 4).

The search and review process outlined above was considered complete when a sense of “saturation” occurred, such that no new themes, approaches, or data were

identified. While there was some diversity in public health topics or issues that were the focus of these research efforts, influenza was by far the most common disease and was the focus of seventeen articles. Of these influenza or influenza-like illness (ILI) articles, most (ten) of the articles were focused on the recent 2009-2010 H1N1 pandemic which may have resulted in both more research and more articles being accepted by journals. Seven articles looked at multiple infectious diseases occurring simultaneously.

The review that follows describes the findings of the literature broken down into groups based on the type of content that was analyzed: news articles (nineteen), Internet search utilization (eighteen), blog/microblog postings (three), and a smartphone application data (one). This breakdown of categories for summarizing results was selected because there are similarities in methodologies that exist in the utilization of the same type of information from the same information stream.

News Article Results

Utilizing newspaper articles (both print and online) was the most commonly reported approach to gathering information from nontraditional sources. Table 4.2 summarizes the research articles identified; the articles are described in further detail below. No consistent health topic was studied; however, topics range from more mundane public health challenges like bedbugs (24) to more violent outcomes like burns (63) and drownings (64). The research utilized data from the past 15 years, with two exceptions: a retrospective look at influenza in Hong Kong (65) and the work of Jensen on cancer news coverage (66). These exceptions may be due to the exclusion criteria of the literature review (only using publications indexed or appearing in 2000 through April 2011), or it may be reflective of the increasing capabilities of computers to facilitate the search, distribution, and identification of news sources.

Table 4.2. *Articles about news media findings*

Primary Author	Title	Study Timing	Outcome of Interest	Source Data	Study Size (# of articles)	Quality
Adelman, RC	Death Makes News: The Social Impact of Disease on Newspaper Coverage	1977-1997	cancer, heart disease, AIDS, diabetes, Alzheimer's, arthritis	CDC & NCI reports	~44,000	Moderate
Anderson, AL	Bedbug infestations in the news: a picture of an emerging public health problem in the United States	2001-2006	bedbugs	None	~370	Poor
Barss, P	Drowning in a high-income developing country in the middle east: newspapers as an essential resource for injury surveillance	1998-2002	drowning	MoH reports	79	High
Boak, MB	Internet Death Notices as a Novel Source of Mortality Surveillance Data	1998-2001	death	state health records	48,651	Poor
Brownstein, JS	Surveillance Sans Frontieres: Internet-Based Emerging Infectious Disease	2006-2007	ID outbreaks (141)	None	n/a	Poor
Brownstein, JS	Information technology and global surveillance of cases of 2009 H1N1 influenza	2009	flu (H1N1)	WHO & CDC	87,000	Poor
Brownstein, JS	Evaluation of Online Media Reports for Global Infectious Disease Intelligence	2006-2007	ID outbreaks	WHO	11,194	Poor
Chan, EH	Global capacity for emerging infectious disease detection	1996-2009	ID outbreaks	WHO	276	High
Collier, N	What's unusual in online disease outbreak news?	2008-2009	ID outbreaks	ProMed	7,482	High
Connor, SM	Newspaper Framing of Fatal Motor Vehicle Crashes in Four Midwestern Cities in the United States, 1999-2000	1999-2000	car crashes	FARS	368	High
Ghaffar, A	Newspaper Reports as a Source for Injury Data in Developing Countries	1999	injuries & fatalities	police reports	438	High
Habel, MA	The HPV vaccine: a content analysis of online news stories	2006	HPV vaccine	ACIP	250	Moderate
Jensen, JD	Making sense of cancer news coverage trends: a comparison of three comprehensive content analyses	2003	cancer	reviewed publications	5,327	Moderate
La Bash, HA	Deployment stressors of the Iraq War: insights from the mainstream media	2003-2005	war-related stress	None	336	Moderate
Nasrullah, M	Newspaper reports: a source of surveillance for burns among women in Pakistan	2004-2005	burns-set on fire	ME records	222	Poor
Nelson, NP	Event-based biosurveillance of respiratory disease in Mexico, 2007-2009: connection to the 2009 influenza A(H1N1) pandemic?	2007-2009	flu (H1N1)	CDC	>2M	Moderate
Reilly, AR	Indications and Warning of Pandemic Influenza Compared to Seasonal Influenza	1967-1968	flu	WHO	n/a	Poor
Takahashi, Y	Analysis of news of the Japanese asbestos panic: a supposedly resolved issue that turned out to be a time bomb	1987-2005	asbestos	None	293	High
Wilson, JM	Media reporting of the emergence of the 1968 influenza pandemic in Hong Kong: implications for modern-day situational awareness	1968	flu	WHO reports	111	Poor

Findings of news media research

News media can provide insights about novel or emerging issues. An increase in both the number of news stories and the severity of cases described in news stories can be seen in news sources from Mexico in the weeks and months leading up to the start of the H1N1 outbreak, which showed that news media provided a prediction of an upcoming reality (17).

There is a lag between newspaper reports and official reporting, with news media often providing the first reports. One research study indicates that, over time, the duration of time between public communication (news) about a disease and official reporting

has decreased from forty days in 1996 to nineteen days in 2009, with unofficial news reporting about a disease being available prior to official reports (67) by an average of twelve days (68). These findings have held for various diseases and have been confirmed using both manual and automated methods for news collection, review, and processing (18). This delay in reporting can be seen as far back as 1968 where the emergence of influenza in Hong Kong was first announced in newspaper stories (65), with newspaper reports of social disruption indicating influenza appearing five weeks before official recognition (69). This faster reporting pace of news compared to official reports can also be seen in death reports where newspaper data provides insight about mortality data up to a week faster than death registries held by state health departments and the CDC (8).

News media can over- or under- represent current health conditions.

Conditions may be overrepresented (breast cancer) as compared to the national incidence rate, while other conditions are underrepresented in the media compared to how often disease occurs (bladder cancer) (66). Media may also focus more often on treatment with limited coverage on preventing, detecting and coping (66). In coverage of injuries caused by vehicular crashes, both the likelihood of restraint use and the risk to drivers was under-represented in newspaper articles as compared to mentions of the involvement of teens in fatal car crashes and the involvement of alcohol in all crashes (70). Trends in newspaper article frequency run parallel to overall mortality trends, but not to prevalence or incidence for cancer, heart disease, AIDS, diabetes, Alzheimer's, and arthritis (7).

News media can provide health information that cannot be verified elsewhere.

Findings that have not been otherwise confirmed were identified using news media; one such example is the reporting of a quintupling of bedbug infestation rates from 2003 to 2006 (24). Additionally, health departments may not disclose findings that are controversial or which may portray a region in a poor (71), but these details can be reliably gained from

newspaper reports of violent acts such as burnings (63, 65) or drownings (64, 67, 72). This inconsistency is also true for acts that a government may not want the public to be aware of such as diseases caused by asbestos exposure (26), or for acts, such as suicide, that have negative cultural implications (73).

Limitations of news media research

Bias is a major limitation of news media research. Writers, editors, and publishers are inherently biased when reviewing the likelihood that any given event will be reported in the news, which may reflect events that affect readership (24). In addition, reporting norms change over time and this introduces another source of news reporting variation (66). For example, the use of scientific terms and disease names may be increasing as these terms become more common amongst the public. Bias may also be introduced into news reporting due to intentional political efforts such as misinformation campaigns and government censorship (26). Finally, reporting can also include language bias since all of the sources used were in English (67), and some studies were focused on specific cities or countries (70, 73). Use of a single media source—newspaper—often limited the scope of findings (66), and these studies could have been extended to include sources such as television or magazines (8).

The ability to access data for verification was limited at times, since organizations like the WHO use private internal websites to share more detailed information than that which is made available to the public (67), and gold standard data, like coroners reports, may not be widely available (64, 67, 72). Additionally, opinions and attitudes about a condition cannot be validated, so there is no way to confirm if findings give an accurate depiction (74).

Quality of news media research

In this literature review, a ‘gold standard’ criterion was met when the accuracy of news media or "crowd sourced" data were compared to official reports from reputable sources such as the WHO, the CDC, ministries of health, or state health departments. This level of comparison occurred for most of these articles (thirteen), but with a few notable exceptions. Three articles did not compare news media results to any type of official data, which leaves those results as questionable since there is no form of comparison between the results and the expected outcomes. For each of these articles (9, 24, 75), there are data sources that could have been used for comparison but were not. This omission would be similar to undertaking a study looking at the natural history of disease while ignoring other attempts to describe the disease in detail.

Studies were considered poor quality if they did not include a description of data extraction (10) or a description of limitations (63, 65). Studies were also considered poor quality if they did not use source data for comparison and were missing extraction or limitation details (9, 24). Studies that were high quality provided descriptions of data methods, analysis approaches and limitations of the work (64, 67, 70, 72).

The studies used various approaches to identify those news articles that would be included in the research, ranging from "manual" review of each article to more automated, computerized approaches. Of those approaches where non-automated methods were used to extract data from news articles, there were three approaches identified. In one method, the research staff developed an extraction sheet or worksheet, which contained explicit variables to be gathered from articles that met inclusion criteria. The number of variables ranged from three (67) to forty (70), based on the outcome being addressed. The second method for manual review was to determine which articles were to be included and then to

assess them based on themes (74) or topics (75). At times, this method relied on multiple staff making independent assessments and then comparing outcomes (73). The third manual method was to include any article where the outcome of interest was included in the title (65) or anywhere in the text (7). For automatic extraction, two findings used methods to extract only the disease and location of interest (9, 72), while another approach involved automated extraction of five variables of interest (8). In the analysis of asbestos news stories in Japan (26), a mix of manual extraction to identify articles and automated extraction to identify concepts and associations of interest was used.

Internet Search Utilization Results

This review identified eleven articles where researchers used Google Insights for Search or Google Flu Trends to make use of crowd sourced data. The exceptions are the work of Hulth, who used a Swedish tool that may be a Google-like analog (14), while Polgreen (30) and Cooper (76) used Yahoo! search terms. All of these studies were conducted between 2004 and 2011, likely because the ability to track and analyze search statistics using the Internet is a relatively new capability. In Table 4.3, it is clear that there is some significant variation in what health issues were studied. While influenza was the most commonly researched health issue (ten), there were articles that described other issues such as chronic conditions (34), and non-infectious disease outcomes such as depression (33), and ophthalmological concerns (77).

Table 4.3. *Articles about internet search findings*

Primary Author	Title	Study Timing	Outcome of Interest	Source Data	Quality
Bentley, RA	A rapid method for assessing social versus independent interest in health issues: a case study of 'bird flu' and 'swine flu'	2005, 2009	flu (AI & H1N1)	None	Poor
Boyle, JR	Prediction and surveillance of influenza epidemics	2009	flu (H1N1)	syndromic	Poor
Breyer, BN	Use of Google in study of noninfectious medical conditions	2005-2010	diabetes, heart attack, high BP	Peer-review publications	Poor
Carniero, HA	Google Trends: a web-based tool for real-time surveillance of disease outbreaks	2004-2009	ID outbreaks (WNV, RSV, AI)	CDC	Poor
Cooper, CP	Cancer Internet Search Activity on a Major Search Enging, United States 2001-2003	2001-2003	cancer	American Cancer Society	High
Ginsberg, J	Detecting Influenza Epidemics Using Search Engine Query Data	2003-2008	flu	CDC	Moderate
Hulth, A	Web Queries as a source for syndromic surveillance	2005-2007	flu	lab & sentinel surv data	High
Leffler, CT	Frequency and seasonal variation of ophthalmology-related internet searches	2004-2008	ophthalmogic concerns	environmental data	High
McCarthy, MJ	Internet monitoring of suicide risk in the population	2004-2007	suicide	CDC	Moderate
Pelat, C	More Diseases Tracked by Using Google Trends	2004-2009	ILI, gastroenteritis, chickenpox	French CDC	Poor
Polgreen, PM	Using Internet searches for influenza surveillance	2004-2008	flu	lab& CDC	High
Reis, BY	Measuring the impact of health policies using Internet search patterns: the case of abortion	2004	abortion	rates, availability	Moderate
Seifter, A	The utility of "Google Trends" for epidemiological research: Lyme disease as an example	2009	Lyme's disease	CDC case reports	Poor
Valdivia, A	Monitoring influenza activity in Europe with Google Flu Trends: comparison with the findings of sentinel physician networks - results for 2009-10	2009-2010	flu	ECDC/WHO reports	Moderate
Valdivia, A	Diseases Tracked by Using Google Trends, Spain	2004-2009	ILI, chickenpox	Spanish CDC	Poor
Wilson, K	Early Detection of Disease Outbreaks Using the Internet	2008	listeriosis	Canada gov't	Poor
Wilson, N	Intepreting Google Flu Trends Data for Pandemic H1N1 Influenza: The New Zealand exp	2009	flu (H1N1)	syndromic	Poor

Findings of internet search utilization research

Internet search patterns align with disease incidence patterns. Search trend data for Lyme Disease matches both the seasonality and the geographic distribution (78). This type of Internet search data also aligns with established patterns for seasonal outbreaks of respiratory syncytial virus (RSV) (20). This phenomenon is not limited to the United States since web queries on a Swedish analog to Google showed that queries aligned with both sentinel and laboratory data for influenza. The Swedish data showed peaks at the same time with a stronger signal during the 2006-2007 flu season, which may have been a result of a more severe influenza season (14). Additionally, search terms in France were highly correlated with incidence of both gastroenteritis and influenza-like illness (ILI) (79).

Internet search patterns align with environmental factors that impact health, especially seasonality. Search utilization can quantify seasonal and environmental variation in eye-related terms (i.e. increased sunlight intensity was associated with increased searches for 'dry eyes' and eye floaters') (77). Searches for depression were also significantly correlated with seasonal temperature variations in both the north and south hemisphere, with the degree of correlation varying by latitude (33). Additionally, search volumes for chronic conditions such as diabetes, high blood pressure, nephrolithiasis, and myocardial infarction aligned with the seasonal morbidity data for each of these conditions (34).

Internet search patterns can both align with disease patterns and provide an early warning or indication of an upcoming increase in incidence. In the United States, Google flu queries were shown consistently to provide estimates of ILI percentage one to two weeks prior to when CDC published information from their sentinel provider network (37). A 2009 comparison in Queensland indicated a high correlation between Google search frequency and emergency department surveillance data. In this case, Google provided earlier indications of increasing demand before case presentations arrived at an emergency department (80). When comparing search trend data to provider visits/consultations for H1N1 in Europe, correlations were higher during the second (fall) peak, with the search terms sometimes identifying the peak one or two weeks prior to the provider data (41), (81). The exception is in Sweden, where search terms indicated the peak eleven weeks prior to the peak seen in the sentinel physician network (an official source) (41). Google Flu Trend data from the first peak of H1N1 in New Zealand aligned with sentinel provider data, except for the week of July 5th where Google showed the incidence peak one week prior to the sentinel network peak (82). For West Nile Virus, Internet search metrics were temporally aligned with the CDC's incidence data. Additionally,

there was an increase in searches for ‘rash’ in the months prior to spikes in West Nile Virus (WNV), which may serve as a proxy or early indicator for the upcoming WNV season (20). During the 2008 listeriosis outbreak in Canada, Internet search terms showed a spike beginning in mid-July, while official reports did not appear until August (83).

Internet search data can be an input parameter for predictions of disease outbreaks. Polgreen and his team used Internet search term data as an input for a model intended to predict the timing of spikes of influenza outbreaks (30). Using data from Yahoo! search, his team predicted influenza rate increases 1-3 weeks prior to laboratory results (30). Internet search patterns can also be indicative of social learning and group direction, as evidenced by the work of Bentley who looked at individual interest in ‘swine flu’ and ‘bird flu’ (84).

Internet search patterns can indicate sub-population variation in outcomes of interest. Search patterns for suicide and self-harm positively correlate with the CDC statistics for youth, but the same correlation did not exist for adults (85).

Internet search patterns can inform people about health issues of concern in places where access to care is an issue. The volume of Internet searches for abortion are inversely proportional to local abortion rates and directly proportional to local restrictions on abortion; therefore, Internet search volume may be a way to assess interest amongst the public about health activities that may be restricted or sensitive (31).

Internet search patterns may provide insight on controversial and often private topics. Studies that looked at more controversial topics such as abortion (31) and suicide (85), suggested that this type of data collection would be extremely useful for these topics since people are less comfortable self-reporting about them.

Internet search patterns do not always align with actual health events. There are also circumstances where search volumes are not indicative of actual disease incidence but are in response to media spikes. This trend was seen with avian influenza in 2005 and 2006, where search frequency spiked in the United States, but no cases were reported (20). The same phenomenon has been seen with increased reporting of cancer stories resulting in increased search utilization (76). New Zealand's HealthLine system received calls for H1N1 that showed a peak three weeks prior to Google Flu Trend data, which may be indicative of variation between Internet users and those without computer access (82). Google Flu Trend data also did not align with news media reports for H1N1; the massive media peak did not relate to increases in disease incidence in New Zealand, but rather to the global concerns about the coming flu pandemic (82).

Internet search patterns may lag behind actual disease patterns. In both Spain (81) and France (79), Internet searches for chicken pox lag approximately one week behind incidence reported by official surveillance systems.

Limitations of internet search utilization research

Internet access and utilization of internet search tools varies across demographic, socioeconomic and geographic population characteristics (31) and information from a search engine company does not provide these user characteristics (85). These data have the potential for a non-representative sampling bias (78), especially due to the overrepresentation of younger and young adult computer and smartphone users (85), since Google searches are less likely to be completed by people under the age of 10 or over the age of 70 (80). Although there is an association between the proportion of the population that used the Internet for health information in Europe and between H1N1 searches and provider visits (41), it may not be sufficient to validate the correlation.

Data were gathered, censored, and selected by the search company owner at a weekly aggregate (77) and the search company (Google) did not provide visibility on mathematical search assumptions and approximations, which may obscure true trends (78). It is unclear if these approaches vary between vendors (76). Additional work is necessary to find suitable Internet search query proxies to be correlated with reported diseases of interest (20), and there is limited evidence on how much of a difference from the baseline indicates an actual outbreak signal (83).

It is not evident that there is a causal effect between someone searching on the Internet and someone actually experiencing the outcome of interest, and we must be careful not to draw conclusions out of a coincidence in data (58). This effect is seen clearly in suicide search data where a person may be looking for information on suicides because they intend to commit the act or because they intend to prevent the act (85). To better assess causality, observational studies must be conducted on the topic of interest to validate findings (31). Internet search results may also be impacted by news events, cultural differences, alcohol consumption or other factors (33) and correlations may be meaningful only across large populations (37).

Quality of Internet Search Utilization

When determining what source data to compare with Internet search data, the most common choice was traditional biosurveillance data (either syndromic or laboratory-based), collected from either the CDC or similar government authorities or from peer-reviewed publications (34). Only Bentley's work on bird flu and swine flu failed to use an official data point for comparison (84), resulting in poor quality based on assessment criteria defined above.

Most of the research (ten articles) using Internet search terms is of poor quality due to the lack of description of specific search terms used, how those terms were validated, and what selections were made when gathering data from Google Insights or other similar tools. High quality research included methodologies for control searches (77), as well as detailed descriptions of search terms and combinations used (76, 86). Polgreen's work provided detailed descriptions of statistics used to analyze the data and details of models developed to predict influenza rates using search data (30).

Blog and Microblog Posting Results

One article was identified that gathered information posted to blogs, and the research specifically looked at postings related to human papilloma virus (HPV) vaccine (87). The authors found that blog postings about HPV vaccines were temporally aligned when there was an increase in mainstream media activity, media controversy, or releases of scientific studies about sexually transmitted diseases (STDs) in teenage girls. Data were extracted from each blog posting via a manual process in which the researchers gathered data on pre-identified variables that included user profile, blog content, size of the bloggers network, and responses of those who used the blog (comments, kudos, replies). The research included blog postings written between November 2005 and May 2008 and excluded postings that were about the virus itself. A significant limitation of this research was that only one social network site (MySpace) was utilized to gather the information, which may limit the generalizability of the findings to users of other blogs or social networking sites. The authors identified a variation between the blogging population and the general population as a limitation that was not incorporated into this analysis (87). Another significant limitation is that this study looked at attitudes and beliefs, and it used news media as its source of comparison, leaving one to question the validity of both the comparison data

(news) and the gathered data (blogs); this resulted in a "moderate" quality rating in my assessment.

Two articles addressed the use of H1N1 information on Twitter, but they looked at two different sub-topics: personal experience with H1N1 (23) and antibiotic misuse (39). For the first analysis, a combination of manual and automatic coding was used to collect data from Tweets posted from May to December 2009, with each Tweet then categorized as personal experience, joke, concern, expression of relief, etc. When frequency of Tweets about H1N1 personal experiences were compared to weekly US H1N1 rates as provided by the WHO, there was a high concurrence (23). The most significant limitation identified is the lack of a well-defined study population, as those who Tweet about H1N1 may not be representative of the entire population of Twitter users, or representative of the general population (23). An additional limitation not addressed in the article is the lack of validation of terms used on Twitter since people have adopted extreme shorthand when working within the defined character limits for posting to the site, and those terms have not been fully evaluated as compared to traditional terminology and language.

The second article that researched the use of Twitter was the work of Scanfeld that looked at evidence of misunderstanding or misuse of antibiotics between March and July 2009 (39). Tweets were manually reviewed and classified into one of eleven categories, with 100 Tweets randomly selected from each category to be used for analysis. In addition to capturing the Tweet text, researchers collected data about the number of followers, the number followed, and the number of status updates by that user. No source data was used to compare the Tweet information to that of an alternative source, so this resulted in a moderate quality finding since the study provided a descriptive assessment with no validation or verification. The most significant limitation is the unclear validity of the content provided, and some of the postings may include embellishment or exaggeration which was

not measurable. Additionally, the novel experience of H1N1 and the additional media coverage may have led to an increased amount of misinformation as compared to a normal influenza season which cannot be validated (39).

Smartphone Application Results

Only one article was identified that used a smartphone application for gathering health information (16). This article described the data collected from inputs to an application titled 'Outbreaks Near Me' which enabled users to provide information about their knowledge and experience related to disease, with 95% of the postings relating to influenza (16). Information submitted via the application was compared to news media reports available on HealthMap and to CDC metrics for sentinel influenza surveillance. Information provided by the general public provided insight that was not otherwise available, especially related to school closure. However, no quantitative data indicating how often this novel information was gathered had been described in the article, so this resulted in a poor quality rating. The most significant limitation was the inability at this time to verify or corroborate data submitted by users (16), meaning that the data collected currently has limited utility for public health officials. While some technologies in development provide an assessment on credibility of the submitter (based on past performance), that approach was not yet utilized for 'Outbreaks Near Me,' and the risk remains of this technology facilitating the spreading of rumors and misinformation (16). A limitation that was absent from the article was the selection bias that exists in the data utilized, since smartphone ownership was not as common across the population at the time of publication (2010) as compared to ownership in 2013.

Discussion

Overall, the review indicated that further research on the utility of news and social information streams is important to continue identifying potential applications for these content sources. The review also suggested that there is a long way to go in this field towards developing a more concrete understanding of both the benefits and the risks of using these sources to gather health information and to establish methodologies and best practices for this field of research. There were a number of articles reviewed that described the possible usage and potential value of news and social media as sources for public health insight, but the articles did not include any actual analysis and were often theoretical in nature [these were not included in this review]. These theories include the potential utility of alternative sources for policy decisions or for improving disease prevention and control programs (88), as well as the concept of using indications of social disruption (such as change in practices like attending festivals) as indirect markers for potential outbreaks (89).

When looking at the findings from the different sources included in this review, there is some variation between the information streams. For instance, while the news can provide insights about novel or emerging issues, it can also over- or under- represent current health conditions. There is a lag between newspaper reports and official reporting, with news media often providing the first descriptions of disease events. Additionally, news media can provide health information that cannot be verified elsewhere. Internet search patterns by the computer-using population align with disease incidence patterns and environmental factors that impact health. Additionally, this data can provide an early warning or indication of an upcoming increase in disease incidence while also being used as an input parameter for predictions of disease outbreaks by officials. These data can inform people regarding health issues of concern in places where access to care is an issue and may provide insight on controversial and often private topics. Conversely, Internet search

patterns do not always align with actual health events and lag behind actual disease patterns. Microblog posts, such as Twitter, provide insight that aligns with official reporting, while a smartphone application provides novel information not otherwise identified, specifically as related to the impacts of H1N1 on schools.

Although these studies have shown value as potential sources of health information from the public, some limitations were identified in the studies. One of these potential areas of concern is the possibility of publication bias, since there were no identified published articles that did not have results that showed no value from news or social media sources. This may be the result of people only doing research on topics for which there is a strong association between media and data. For the findings related to newspaper articles, there are limitations about the applicability of these findings across varying geographies and types of newspapers around the country or the world. News producing entities can be categorized as local (for a specific city or region) or national, with additional variation between organizations that are affiliates of bigger national news outlets and those organizations that are completely independent. Additionally, some online news services no longer have a paper edition and are Internet-based only. This impacts content, because when comparing print to online for American news content, there is an increase in opinions and light news online while informational news is less present online as compared to in print (90). News sites are sometimes the partners of radio and television stations (either local stations or affiliates of national networks).

Additionally, for the three studies that looked at social media sites and applications for smartphones, there is an inherent limitation to generalizability in these findings because there may not be comparability between those who can both afford and choose to own smartphone devices compared to those who do not. Due to varying use patterns of computers, the Internet, and search engines, there is also a limit to generalizability, since we

cannot confidently apply search pattern findings for people who have access to the Internet and use it for medical information of the general population.

When reviewing the articles for quality, it was apparent that there is no standard or consistent methodology for assessing the quality, validity, or accuracy of the information gathered from news and social media sources. There is also no standard approach or methodology for comparing information from news or social media sources to formal reports and official sources. Some researchers took the initiative to look for correlations between the data they identified in social or news sources and official reports from health agencies or government organizations, but many researchers did not do any such comparison. When assessing quality, a particular reality was emphasized: it is hard (or impossible) to compare quality across different media (newspaper to Twitter) as each has its own unique shortcomings and strengths.

The most significant limitation of this literature review is that gray literature was not included, which may have resulted in missing articles from academic institutions, foundations, and think tanks. The other significant limitation is that this topic is relatively new (especially as it relates to social media websites and tools), and sources of content are evolving so rapidly that there is not a great deal of research being published in journals at this time. One other limitation of this work is that only articles written in English were included, which may exclude worthwhile efforts in other parts of the world where use of technology is more prevalent. Additionally, this review did not include any classified government or proprietarily sequestered studies. These may have added insight into how certain parts of the government are using these technologies for health monitoring with intelligence-based methodologies.

CHAPTER 5. METHODS

This dissertation has two key outcomes that are being studied. The first is an assessment of the utility of infodemiology data sources (news media, Internet search patterns, and social media) when compared to official epidemiology reporting. That assessment was first done by doing a descriptive analysis of each data source independently, and then comparing each infodemiology source to the NNDSS data set. Then, the three infodemiology sources are combined to assess the ability of news media, internet search, and social media to serve as a signal or indicator in advance of official reporting. The second key outcome was to assess the knowledge and attitudes amongst disease surveillance experts on if (or when) they would use infodemiology sources to inform their work. The survey was necessary to understand existing opinions on novel sources and to understand the likelihood of these sources being incorporated into public health practice (regardless of the outcomes on the utility assessment). Even if the infodemiology sources show utility, if practitioners are not willing to use the new information, this line of data is not worth further investigation.

Methods for Data Collection and Processing

For this research, data was collected from each of the infodemiology sources including: news media, internet search frequency, and social media. For the "gold standard" or "ground truth" data to which all other data were compared, the National Notifiable Disease Surveillance System (NNDSS) from the Centers for Disease Control (CDC) was used. For news articles, content scraped from online news aggregators by HealthMap during the study period was provided. For search term patterns, Google's

Insight for Search data provided relative search frequency. For Twitter, data were pulled from an archive of the Twitter fire hose by Gnip. In addition, original survey data came from the responses of study participants. Each data set (excluding the survey) was collected for the retrospective period from January 1, 2010 to December 31, 2010. A list of terms was generated to include clinical terms as well as common misspellings and Internet slang (23) for each data source to include the scientific genus name ("Bordetella"), as well as the disease name ("pertussis") and the common term ("whooping cough") (see Appendix B). While news articles are expected to have correct spellings, Google search auto-corrects spelling for search terms, while investigations of Twitter must include variations of spelling and verb tense to account for most postings. The list in Appendix B was generated from Google Insights for search results at the national level. The intent was to maximize the number of true positives (genuinely pertinent articles or Tweets) and minimize retrieval of false positives (irrelevant articles or Tweets that are excluded due to alternate meanings of the key word) (7).

Official Reporting

Official reports are epidemiology data that are considered as the referent or “gold standard” for the event. For this research, the provisional case counts from the Center for Disease Control (CDC) National Notifiable Diseases Surveillance System (NNDSS) were used for all weeks in 2010. NNDSS data are reported by states and territories weekly, and published as part of the *Morbidity and Mortality Weekly Report* (*MMWR*) (91). These case counts are considered provisional for 2010 due to both ongoing revision of information by state health departments as well as delayed reporting (92), so later weeks in the series may reflect changes in earlier weeks as additional cases are identified (91). Crude values were used (rather than age-adjusted), because those rates reflected population-wide impact (7).

The researcher used the CDC's MMWR Tables web interface (91) to select the year of interest (2010) and the table of interest (Table II, Part 7) for each week of the year. Numbers were pulled for both "Current Week" values and "Cum 2010" (total for the year) values for each of the three states being studied (California, Michigan, and Ohio). The pulled data were put into state-specific Microsoft Excel spreadsheets. To ensure accuracy of the data (as well as to maintain a record of what the provisional values were), the researcher copied the website's content for every week in 2010 in Microsoft OneNote.

When reviewing the weekly counts against the cumulative counts for each of the states, it was clear that concordance was lacking between the expected cumulative count and the actual cumulative count in the original data tables online. In order to get a more accurate depiction of the weekly count using a retrospective approach, the researcher had to calculate a revised weekly incidence value using the difference between the cumulative for the week and the past week's value. For example, the 2010 NNDSS data for California shows a cumulative case count (prevalence) of four cases at Week 9. Then, in Week 10, two new (incident) cases are reported and the cumulative case count is seven. One would expect from arithmetic that if there were four total cases in a set time period (through Week 9) and two additional cases in the following week (Week 10), that the new total case count at the end of Week 10 would be Week 9's total [4] added to Week 10's new cases [2] to result in six total cases at the end of Week 10. That is not how the data in NNDSS are presented. This discrepancy between the past total cases plus new cases not equaling the new total cases resulted in the researcher needing to determine to which week to allocate the yet unlisted cases. The researcher decided to use the total case count per week to calculate a revised weekly count (by subtracting Week 9 from Week 10, the case count for Week 10 was determined to be three). This approach linked the revised weekly case count (the calculated difference in numbers) each week to the previous week, which may or may

not be accurate if the cases were reported a week late or many weeks later. Lastly, the researcher identified each instance where the cumulative value did not align with the expected value and performed arithmetic checks on all calculations.

In the NNDSS data, there was one week where the values for cumulative count decreased from the previous week (Week 32) for both California's data and nationwide data. There was also one week (Week 38) for Ohio where the cumulative count decreased. These negative values are likely due to a calculation or reporting error in previous weeks reporting being corrected on that single week. To eliminate the appearance of a decrease in cumulative count, the researcher considered those weeks as no data. When (and if) the provisional counts from NNDSS are updated for 2010, the updated data could be included for analysis.

News Media

As part of their normal operations, the HealthMap system utilizes automated querying and filtering of web-based reports of infectious disease, displaying an average of 1,000 alerts at any given time (28). The system's accuracy in automated categorization of disease and location is described as 81% to 91% concurrence, and analysts reviewed many articles manually to confirm categorization accuracy (28). In 2010, HealthMap pulled English language web news articles via three mechanisms: a) articles that *Moreover* (a media monitoring service) identified as public health-related, b) articles from predetermined Google searches and, c) articles submitted by HealthMap staff and/or community users (93). These articles can be sourced back to online news wires (including Google News), really simple syndication (RSS) feeds, and expert-curated accounts like ProMED Mail (29). An automated technology created by HealthMap groups articles into clusters; this grouping is especially important, because many articles use the same Associated Press story as a source (93).

For this research, the subset of articles from 2010 was provided in two data sets as text files. The first data set included articles with a defined location in California, Michigan, or Ohio. The second set included alerts tagged as United States, to capture any articles with reference to multiple states. If either “pertussis” or “whooping cough” was in the article (to include either title or body) and was not solely about policy or did not describe incidence, the article was tagged for inclusion in the HealthMap system. HealthMap provided the first publication of any articles to reduce the number of duplicates needed for review.

HealthMap provided data sets as text files, which were then imported into Microsoft Excel. Of the two data sets, the first [“Just_US”] included thirty-one news articles that reference both pertussis and multiple states or the United States broadly (28). The other data set [“3 states”] included ninety-eight news stories that referenced both pertussis, and California, Michigan, and/or Ohio (28). The articles were then reviewed to ensure that the automated processes worked as expected. Within the “Just_US” data set, four news stories included information regarding California, Michigan, and/or Ohio. Within the “3 states” data set, twelve of the articles were duplicates, five articles were not found (due to time since original posting), and three links were to podcasts or videos which were not the focus of this research. As a result of these exclusion criteria, seventy-eight articles from this data set remained to be included into the research.

The researcher then completed the extraction of key information manually. Each relevant article link was archived for future use and analysis along with key elements of information extracted from the articles for further analysis. The extracted content included location, date of article, source, case count, fatality count, and any other epidemiologic details (to include incidence and prevalence). If additional news articles were linked as part of the original story, they were reviewed for potential inclusion. Articles were classified as ‘novel’ or not if they provided the first report of any new information (cases, location,

deaths). If a date was referenced in the article itself, that date was not used. The publication date of the article was used since that is the date the information would have been made available to public health officials.

Google Search Terms

Google Insights for Search provides relative search volume and probability for specified time and geographic regions (34). More specifically, Google provides the likelihood that a random user will search for the terms of interest, from a specific location, at a specific time (34). Raw search volume is not publicly available from Google, which reduces the impact that population size and Internet prevalence would have on detecting changes or trends. Output rates are normalized on a scale of 0 to 100 (34), and they are available in both a relative and fixed scale (94). Google Trends (GT) analyzes a fraction of the complete Google web searches over a period of time and extrapolates the data to estimate search volume, and updates this information daily (20). To determine the geospatial information within GT, the user's Internet protocol (IP) address is utilized and this provides a rough identification of the source location for the informant (20).

Over time, there has been an increase in the volume of web searches as Internet access becomes more available. Because of this, the average volume increases over time and so the average search denominator (total searches) continues to increase over time (20). As a result of these ever-increasing values, the sensitivity in detecting changes in future search volume continues to decrease (20). GT controls for this problem by using an unrelated common web search query. This normalization also compensates for population size which makes it possible to rank cities based on search volume trends (20).

Using the Google Insights web interface (95), the researcher pulled normalized relative search volume (96, 97) for "pertussis" OR "whooping cough" and downloaded

results as a comma-separated values file (.csv) in Microsoft Excel for each of the three states as well as nationwide. To ensure accuracy of the data pull (as well as to maintain a record of what the provisional values were), the researcher copied the website's content into Microsoft OneNote. This data pull included the top cities within the state (if available), as well as top searches for the terms. Each "Top search" indicated by Google Insights as potentially related to the search terms used was reviewed to see if it included additional terms that could be utilized, excluding searches for vaccines alone (since relationship or causality could not be verified). For California, the only new term was bordetella (the bacteria that causes pertussis), which was likely searched due to concerns for "kennel cough" and therefore, was dismissed for this research. The same process was also followed through for obtaining data about searches for "pertussis" and "whooping cough" independently, although those results will not be used for this work. The search frequencies for either term independently would only give us a portion of insight about the whole population's desire to find out about pertussis. Since the intent of this work is to look at crowdsourcing, we cannot assume that everyone that searches in Google will use either the clinical name or the common name and must incorporate both terms in our assessment. The intent was to harvest the wisdom of the crowd via existing tools and technologies which includes individuals with mixed backgrounds and terminology uses.

Social Media

A data set containing 70,399 Tweets was provided by Gnip, a company with access to the full Twitter fire hose focused on making social media data available (98). Only Tweets that included one of the pre-defined whooping cough related terms (see Appendix B), and were from profiles that included reference to CA, OH, or MI in their location, and were posted from between January 1 to December 31, 2010 were included in the data set provided. Each Tweet was reviewed and classified as either relevant to pertussis or

whooping cough [n=41], not relevant to pertussis or whooping cough [n=70,278], or excluded because the Tweet was not in English or was all symbols rather than text [n=80].

Tweets were then categorized following Vieweg's classification scheme of on topic and relevant to situational awareness (R), or on topic and not relevant to situational awareness (N) (47), with an added category of prevention (P) to differentiate items related to vaccines or public information campaigns, as well as a category for Tweets that link to news articles (A). Tweets were also reviewed to determine if content written was about the poster or another individual. For any Tweets containing a link to a news article, the researcher attempted to open the embedded link and reviewed any content on the linked page. Following this assessment, the Tweets were assessed for credibility using the methodology developed by Gupta (46). Each Tweet deemed relevant was categorized as “Definitely Credible,” “Seems Credible,” “Definitely Incredible,” or “Can't Decide” by using the following definition of credibility: "the quality of being trusted and believed in" (46). Lastly, a final data set was created where extracted information on location, date of article, and relation (self or other person) was included, along with the full original text.

Survey of Surveillance Professionals

A computerized, self-administered questionnaire (see Appendix C) was developed using UNC’s Qualtrics Research Suite to gather responses to an anonymous survey assessing likelihood of public health decision makers using various types of infodemiology data. Most questions had quantifiable responses using Likert scales only, except for the last two questions that were free form text. A grid format was used as often as feasible to eliminate redundancy, make the survey appear shorter, and require less effort from the participant – each of which may potentially contribute to improved data quality (99). A progress indicator was used to show the respondent their level of completion on each screen as a way to reduce the dropout rate (99).

The survey was distributed to the membership email list of the International Society for Disease Surveillance (ISDS), whose "400+ membership represents professional and academic subject matter experts in the fields of public health surveillance, clinical practice, health informatics, health policy, and other areas related to national and global health surveillance" (100).

Methods for Data Analysis

Data Exploration

The first step in the analysis was to complete a descriptive overview to determine if the raw data yielded any general patterns or substantively interesting findings (7). The data type varied by source: case counts (for NNDSS data), articles (for news), relative frequency (for search terms), and mentions (on Twitter), were calculated for each of the three states per week. These data points were captured for each available date, recognizing that there were weeks for which no data were available from these sources over the course of the study period. Minimums and maximums were identified for each predetermined data types for the selected state (see Table 5.1) and each data source. Next, the statistical mean, standard deviation, and lower and upper 95% confidence limit were calculated within SAS (101). For categorization, any terms that were identified in the search as being related to whooping cough will be combined when determining values (i.e. "whooping cough" and “pertussis” search term frequencies were looked at as a joint value). For the only data set where there was a value per week (internet search frequency), PROC UNIVARIATE was calculated within SAS (101) to obtain the Student's t-test value to assess the distribution of the values over the 52 weeks.

Table 5.1. *Variables assessed per data source*

	Official Reports	News Reports	Tweets	Search Volume
Location	X	X	X	X
Confirmed Cases	X	X		
Suspected Cases		X	X	
Date of Report	X	X	X	
Tweet Frequency (referencing self)			X	
Tweet Frequency (referencing others)			X	
Relative Search Volume				X

Actual peak values (maximums) per state were identified, as well as values that fell one (or more) standard deviation(s) above the mean per data set. Values greater than one standard deviation of the average for that state for 2010 were investigated for a potential tie to a news story or some other external factor. Values greater than three standard deviations from the mean were identified for potential exclusion, since they were likely the result of nonmedical events as seen in previous literature looking at search terms (34). Epidemic curves were generated per data source, with separate lines for each of California, Michigan, and Ohio.

For the analysis of the survey responses, the data were made available as a dataset with categorical values assigned to each response. A summary report was also developed directly from Qualtrics, the system used to contact and collect responses. Qualtrics is an on-line survey tool that is licensed to academic institutions and commercial users (https://www.qualtrics.com). A descriptive analysis of findings was done to look at each question individually.

For the two open-ended survey questions, each response was reviewed and if multiple concepts were in a single response those concepts were treated individually. Upon reviewing all responses, themes were identified and each response was categorized under

one of these themes. Each set of responses within a theme was reviewed to identify key characteristics of the response, and these are the points that are included in the results. For responses that were unique and did not fall into a category, those responses were not used for this research since the intent was to find common or major themes.

Comparative Analysis

The data were examined separately for each of the states, and included assessments for three characteristics: timeliness, accuracy, and correlation.

a) Timeliness was assessed by comparing the timeliness of sources to the official reports (looking for the first reports of a significant increase). This assessment was first done visually by generating an epidemic curve with separate lines representing the official and novel data source. Then, the maximum weekly case counts from NNDSS were compared to the peak value in each novel source. Negative timeliness indicated that the peak in an alternate source preceded the peak in the NNDSS, whereas positive results mean NNDSS data precedes the alternate source (102). This same assessment was conducted a second time to look at all weeks where the value was above one or more standard deviations from the mean (in some cases this was the same as the peak week, but not all). Again, the difference between the first significantly increased (greater than one standard deviation) novel source week and the first significantly increased NNDSS week was calculated.

b) Accuracy was assessed by looking at the ability of the source to discriminate between outbreak and non-outbreak weeks. This assessment was done through sensitivity and specificity calculations, where sensitivity values indicated the ability of the novel data source to identify true epidemic weeks (based on NNDSS), while the specificity value indicates the ability of the novel data source to identify true non-epidemic weeks. A True Positive (TP) was defined as a week where the

infodemiology source indicated an outbreak week, and there was an actual outbreak week in the NNDSS data, while a False Positive (FP) was a week where the infodemiology source indicated an outbreak but the NNDSS data did not. A True Negative (TN) was defined as a week where both the infodemiology source and NNDSS data indicated a non-outbreak week, while a False Negative (FN) was a week where the infodemiology source indicated the week as a non-outbreak but NNDSS data indicated an outbreak.

Sensitivity = TP / (TP + FN)

Specificity = TN / (TN + FP)

Positive and negative predictive values were calculated for each novel source to assess the data's ability to forecast significant increases in case count accurately. A strong positive predictive value indicates that there are a small number of false alarms (72).

Positive Predictive Value (PPV) = TP / (TP + FP)

Negative Predictive Value (NPV) = TN / (TN + FN)

Lastly, the F-1 scores were calculated for each data source within each state to assess overall accuracy [*F1=2 * (PPV * Sensitivity)/ (PPV + Sensitivity)*] (72). This calculation gives equal weight to the probability that a true alert will be found and the probability that a system alert will be a true alert (72).

c) Correlation between the official ("gold standard") data from NNDSS and each infodemiology source was assessed by calculating Pearson's correlation coefficient (16), (79), (83) within SAS (101) to determine if the novel information sources were able to either accurately predict or match official data (specifically change in weekly incidence above expected values). Pearson's correlation coefficient will provide a value indicating the strength of association between the NNDSS data with each of the infodemiology sources (103). A perfect direct association exists if the value for

the correlation coefficient is +1, while a perfect indirect association exists if the value is -1 (103). For these calculations, the researcher ran PROC CORR within SAS (101) to obtain the coefficient and the associated p-value.

CHAPTER 6. RESULTS

This chapter will describe the results of the methodology described previously. The chapter begins with a description of the official NNDSS data and provides insight into the data each infodemiology source was compared to. Then, each source (news media, internet search, and social media) is described individually followed by a comparison of each source to the official NNDSS data. Lastly, the chapter concludes with a description of the results from combining all the infodemiology content and from comparing that fused data set against official NNDSS data.

Descriptive Analysis of NNDSS Data

For the NNDSS data, the first analytical step was to look at the descriptive statistics output from SAS (101), as well as minimums and maximums (Table 6.1). Following that, the data were displayed to reveal the epidemic curve for each of the three states, where the total cases reported per week were graphed over time (Figure 6.1). The researcher also looked temporally at the weekly case counts and compared the weekly case counts to the mean per state (Figure 6.2).

Table 6.1. *Summary of findings from 2010 NNDSS tables (91)*

State	Total Cases	Weekly Mean	Std. Dev.	Lower 95% C.L.	Upper 95% C.L.	Weekly Max.	Max. Week	Weekly Min.	Min. Week
California	3,080	75.18	168.92	27.67	122.69	1,155	31	0	1, 2, 6, 8, 20, 21, 30
Michigan	1,500	28.85	18.28	23.76	33.93	75	44	0	16 & 43
Ohio	1,806	35.51	19.06	30.15	40.87	80	48	8	18

Figure 6.1. *Total pertussis case counts were used to create an epidemic curve per state, based on NNDSS 2010 data.*

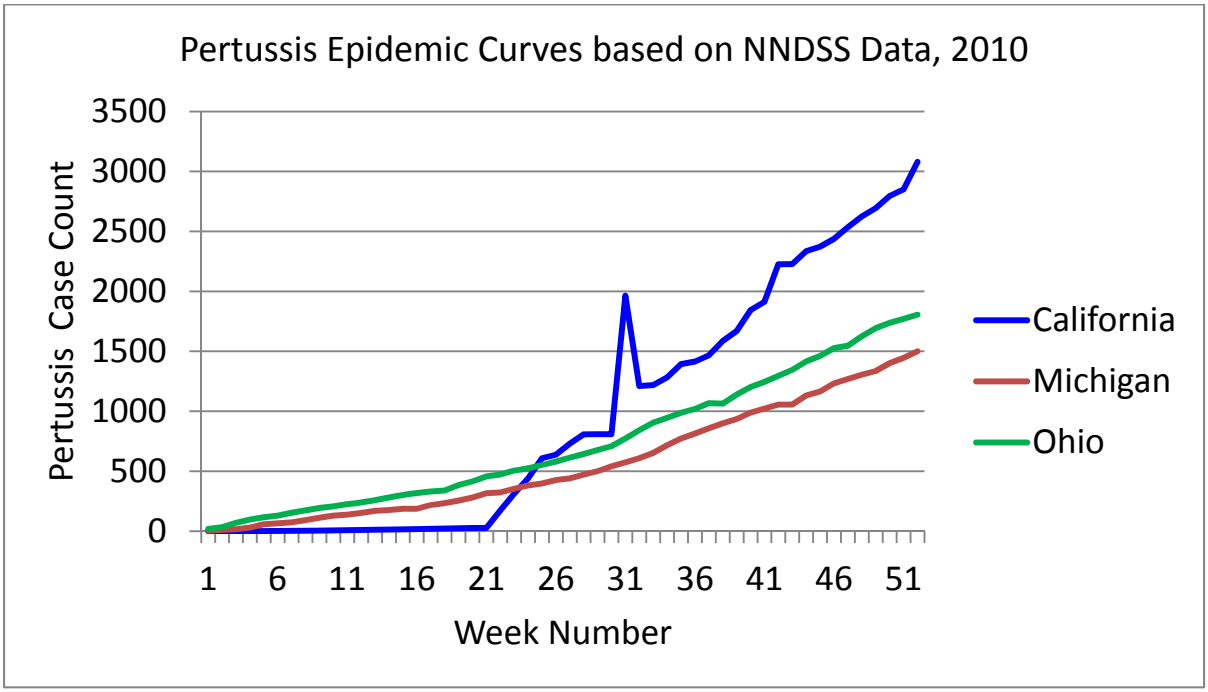
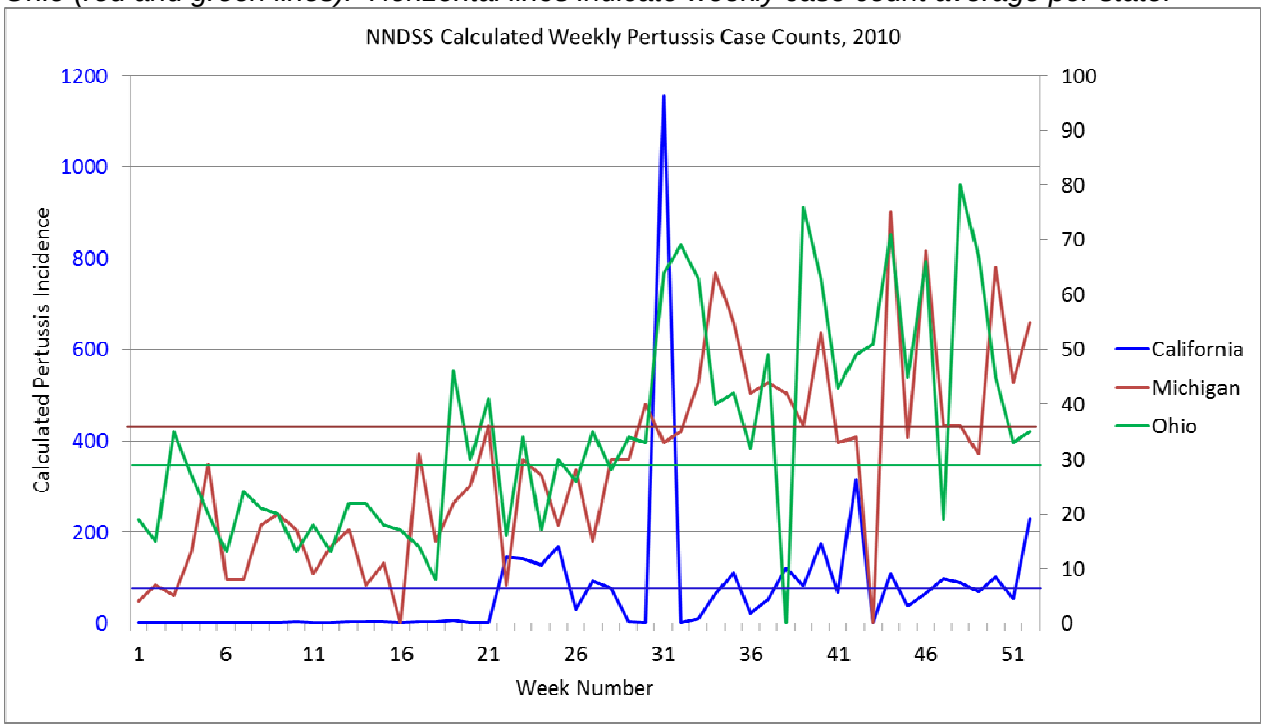


Figure 6.2. *2010 weekly pertussis cases in California, Michigan, and Ohio from NNDSS. Left axis shows count for California (blue line); right axis shows values for Michigan and Ohio (red and green lines). Horizontal lines indicate weekly case count average per state.*



The key comparative statistic was the arithmetic mean number of cases reported in any given week over the one year period. This was a relevant comparator as there were cases reported in almost all weeks for two states (Michigan and Ohio), with California having a twenty-one week period early in the year with no cases reported. In addition to looking at peak weeks to assess increased case counts, the weekly case counts were reviewed to identify weeks that were greater than one standard deviation and greater than three standard deviations above the mean (Table 6.1 and 6.2). For California, there were two weeks falling statistically above the expected incidence, and therefore, these could be considered epidemic weeks for the NNDSS data. Week 31 was three standard deviations above expected incidence, and Week 42 was one standard deviation above expected. Week 31 was also the actual (raw case count) epidemic peak. In both Michigan and Ohio, there were a number of weeks with case counts over one standard deviation away from the mean, but no weeks where the count was three standard deviations higher. For both states, the actual (raw case count) peak was over one standard deviation above the mean; however, there were weeks that were only one standard deviation above the mean prior to the peak week (three times for Michigan, seven times for Ohio). The peak weeks for each state are indicated in red in Table 6.2.

Table 6.2. *Tabulated weekly case counts using cumulative weekly counts from 2010 NNDSS tables (91) with weeks over one standard deviation higher in bold and over three standard deviations in italic bold. The peak week for each state is indicated by red font.*

Week #	End Date	California	Michigan	Ohio
1	1/9/2010	0	4	19
2	1/16/2010	0	7	15
3	1/23/2010	0	5	35
4	1/30/2010	1	13	27
5	2/6/2010	1	29	20
6	2/13/2010	0	8	13
7	2/20/2010	1	8	24
8	2/27/2010	0	18	21
9	3/6/2010	1	20	20
10	3/13/2010	3	17	13
11	3/20/2010	1	9	18
12	3/27/2010	1	14	13
13	4/3/2010	2	17	22
14	4/10/2010	2	7	22
15	4/17/2010	3	11	18
16	4/24/2010	1	0	17
17	5/1/2010	2	31	14
18	5/8/2010	2	15	8
19	5/15/2010	6	22	46
20	5/22/2010	0	25	30
21	5/29/2010	0	36	41
22	6/5/2010	145	7	16
23	6/12/2010	140	30	34
24	6/19/2012	128	27	17
25	6/26/2010	168	18	30
26	7/3/2010	30	28	26
27	7/10/2010	92	15	35
28	7/17/2010	77	30	28
29	7/24/2010	2	30	34
30	7/31/2010	0	40	33
31	8/7/2010	***1155***	33	**64**
32	8/14/2010	.	35	**69**
33	8/21/2010	9	44	**63**
34	8/28/2010	64	**64**	40
35	9/4/2010	110	**55**	42
36	9/11/2010	21	42	32
37	9/18/2010	52	44	49
38	9/25/2010	121	42	.
39	10/2/2010	83	36	**76**
40	10/9/2010	174	**53**	**63**
41	10/16/2010	68	33	43
42	10/23/2010	**314**	34	49
43	10/30/2010	1	0	51
44	11/6/2010	108	**75**	**71**
45	11/13/2010	37	34	45
46	11/20/2010	66	**68**	**66**
47	11/27/2010	98	36	19
48	12/4/2010	89	36	**80**
49	12/11/2010	69	31	**67**
50	12/18/2010	102	**65**	45
51	12/25/2010	54	44	33
52	1/1/2011	230	**55**	35

Descriptive Analysis of News Articles

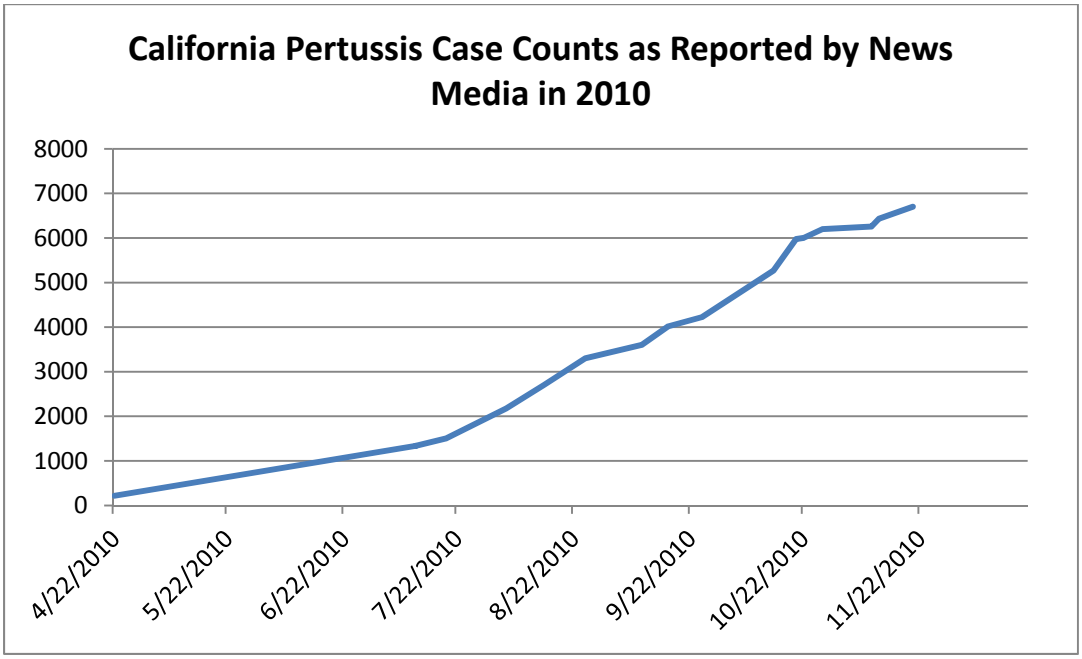
In reviewing and extracting data from the HealthMap-provided news articles (28), articles that provided details about cities or counties were not utilized in this research. Since the "gold standard" (comparison) data from NNDSS is state level, all the corresponding data from other sources was analyzed at the state level. For the extracted HealthMap articles, the articles were sorted by date and novel articles (reporting a number not yet reported) were identified. This extraction resulted in seventeen articles about California, four about Michigan, and three about Ohio (Table 6.3 – 6.5). Since the reporting in California was sufficient to do so, the researcher created an epidemic curve for pertussis using only the news reports for the data points to assess case counts (Figure 6.2). There was one date (8/25/10) where two different reports were in the news with two different estimates of cases; therefore, the larger number was used to generate the epidemic curve.

Table 6.3. *Novel news articles regarding pertussis in California from HealthMap (28)*

Date	***Cases***	***Source***	***Title***
4/22/10	219	NBC - Bay Area	*Whooping Cough Outbreak in California Kills Two*
7/11/10	1,337	National Ledger	*Whooping Cough Outbreak - Health Officials Urge Parents to Be Proactive*
7/19/10	~1,500	alipac.us	*Vaccine urged as Whooping Cough Epidemic Grows*
8/4/10	2,174	MedicalNewsToday.com	*Whooping Cough Epidemic Grows - Health Officials Urge Vaccination And Timely Diagnosis*
8/14/10	>2,700	NPR	*Deadly Whooping Cough, Once Wiped Out, Is Back*
8/25/10	3,000	NewScientist	*Whooping Cough Outbreak Could Be Worst in 50 Years*
8/25/10	3,300	Carlsbad Current Argus	*NM Appears to Be Free of Whooping Cough Outbreak*
9/9/10	3,600	San Francisco Examiner	*San Mateo County Fights Whooping Cough with Vaccine*
9/16/10	4,017	CBS2.com	*Whooping Cough Declared an Epidemic in California*
9/25/10	4,223	MedicalNewsToday.com	*4,223 Whooping Cough Cases this Year in California So Far, Highest in 55 Years*
10/14/10	>5,270	MyMotherLode.com	*Whooping Cough Cases on the Rise*
10/20/10	5,978	CNN	*10 Infants Dead in California Whooping Cough Outbreak*
10/22/10	~6,000	Patch - Imperial Beach	*Two Whooping Cough Cases Confirmed in Imperial Beach*
10/27/10	>6,200	Marin Independent	*Whooping Cough Cases Persist in Marin but Outbreak eases*
11/9/10	6,257	Hispanically Speaking	*California's Latino Community Hard-hit by Whooping Cough Outbreak*
11/11/10	6,431	Bell Gardens Sun	*California's Whooping Cough Outbreak Continues*
11/20/10	6,700	CBN	*Whooping Cough Outbreak Spreading Across US*

Of the seventeen articles on pertussis with reference to California, most (eight) were from national news organizations, of which five of the articles were from national internet-based news sites. Six of the articles were from local news organizations, of which four were internet-based only, while two of the articles were from websites with corresponding papers that are distributed in addition to the website. Two of the articles were from local affiliates of national news organizations (NBC and CBS). Lastly, one article was from a self-identified political action committee (ALIPAC) – this article was the only one identified for any of the three states, and this specific organization is not public health focused.

Figure 6.3. *Epidemic curve for pertussis in California in 2010 utilizing news articles from HealthMap (28)*



For Michigan, there were four articles (Table 6.4) within the data set that had reference to both Michigan and pertussis. The first two reports (one of which was significantly earlier than all other news reports in July) did not provide specific numbers, but indicated the case count to be above a reported number. Interestingly, both of those first articles were from radio stations that have an online news sites (one local and one local NPR affiliate), while the later articles were from national internet-based news sites. There was also a distinctive variation in reporting in the two November articles, with the earlier one indicating 400 more cases than the number included in a report five days later; however, that >1,300 report was very similar to an end-of-year number reported by Yahoo! on December 31, 2010.

Table 6.4. *Novel news articles regarding pertussis in Michigan from HealthMap (28)*

Date	***Cases***	***Source***	***Title***
7/13/10	>600	Michigan Radio	*Whooping Cough on the Rise in Michigan*
11/22/10	>1,300	WKZO	*Whooping Cough Cases up in Michigan*
11/27/10	917	huliq.com	*CDC: Whooping Cough Cases Increasing in California, Michigan*
12/31/10	1,305	Yahoo!	*Detroit Suburbs at Heart of Michigan's Whooping Cough Outbreak*

For Ohio, there were only three articles (Table 6.5) within the data set that fell within the given timeframe that had reference to Ohio and pertussis. Only one of these articles had a case count number. Additionally, of the articles, one was from a local affiliate of a national news corporation (ABC), while the other two were from national internet-based news services.

Table 6.5. *Novel news articles regarding pertussis in Ohio from HealthMap (28)*

Date	***Cases***	***Source***	***Title***
8/4/10	-	ABC7Chicago.com	*Experts: Whooping Cough Outbreak Largest in Decades*
8/7/10	-	examiner.com	*Whooping Cough Cases Still Increasing in Upstate New York and Nationally*
12/2/10	1,546	Yahoo.com	*CDC Team Investigating Ohio Whooping Cough Outbreak*

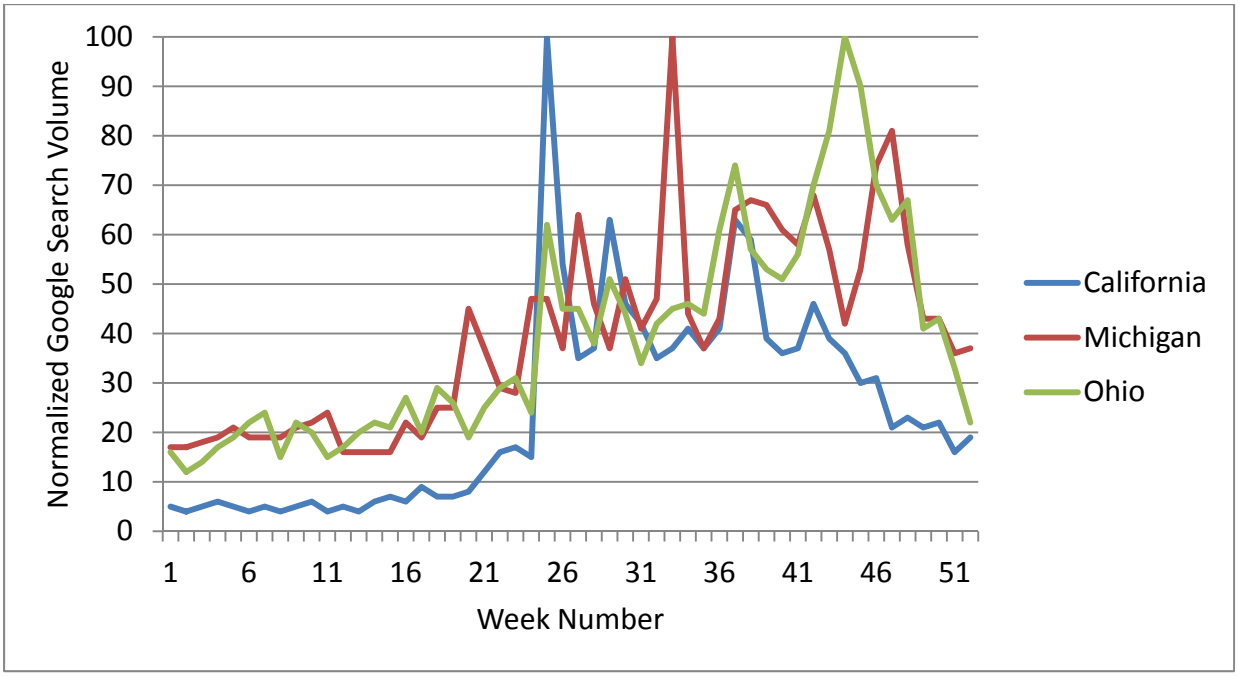
Descriptive Analysis of Google Search Terms

For the Google Insights data, the first analytical step was to look at descriptive statistics output from SAS (101), as well as minimums and maximums (Table 6.6). Following that, the data were visualized temporally to see the epidemic curve for each of the three states (Figure 6.3). Misspellings of prominent search terms emerged in the nationwide search, but these misspellings did not appear in the overall results for any of the three specified states. (The list of terms and misspellings is in Appendix B.)

Table 6.6. *Summary of findings from Google Insight Relative Search Frequency for "pertussis" or "whooping cough" in 2010* (94)

State	**Weekly Mean**	**Std. Dev.**	**Lower 95% C.L.**	**Upper 95% C.L.**	**Weekly Max.**	**Max. Week**	**Weekly Min.**	**Min. Week**
California	24.58	20.67	18.82	30.33	100	25	4	2, 6, 8, 11,14
Michigan	39.62	19.90	34.08	45.16	100	38	16	12, 13, 14, 15
Ohio	*39.12*	*21.51*	*33.13*	*45.10*	*100*	*49*	*12*	*2*

Figure 6.4. *Google Insight Relative Search Frequency for "pertussis" or "whooping cough" cases California, Michigan and Ohio per Week in 2010 (94)*



Google Insights provides users with the top cities where the search terms were used within the selected state. For California, Walnut Creek had reached a relative frequency of 100 (the highest possible value) and San Luis Obispo reached 48. Michigan reached a relative frequency of 100 in Detroit, but no other cities had sufficient search volumes to be included in the results. For Ohio, three cities had sufficient data for search volumes to be available: Columbus (100), Cincinnati (66), and Cleveland (53). When looking at nationwide Google results, the top six cities where pertussis or whooping cough was searched were all in California; however, the city level data were excluded since this research focused on a statewide official reporting source.

In addition to looking at peak weeks to assess increased case counts, the weekly case counts were reviewed to identify weeks that were greater than one standard deviation higher than the mean weekly incidence and greater than three standard deviations (Table 6.7). In both California and Michigan, there was only a single week for each that was three

standard deviations above the mean, and it matched the previously identified raw count peak for each (Week 25 and Week 33, respectively). For Ohio, no weeks had a relative search frequency that fell three standard deviations above the mean. This void may be due to the statistical artificiality of using Google Search, where the defined maximum is 100 while the statistical value was 103.65. In all three states, there were a number of weeks with case counts over one standard deviation away from the mean. For California, none of these single outliers occurred prior to the peak week (Week 25). For both Michigan and Ohio, there were weeks that were one standard deviation above the mean prior to the peak week (one time for Michigan, four times for Ohio).

One additional assessment of the Google search frequency data was done per state to assess the distribution of the values for each of the 52 weeks of the year. Specifically, the Student's T-test was run using SAS (101) resulting in values of 8.57, 15.36, and 13.12 for California, Michigan, and Ohio respectively. For each of the results, the p-value was <.0001 indicating that there is a high likelihood that the patterns of occurrence adhere to an almost normal distribution (103).

Table 6.7. *Weekly data from Google Insight Relative Search Frequency for "pertussis" or "whooping cough" in 2010 (94) with weeks over one standard deviation higher in bold and over three standard deviations in italic bold. The frequency peak week for each state is in red font.*

End Date	California	Michigan	Ohio
1/9/2010	5	17	16
1/16/2010	4	17	12
1/23/2010	5	18	14
1/30/2010	6	19	17
2/6/2010	5	21	19
2/13/2010	4	19	22
2/20/2010	5	19	24
2/27/2010	4	19	15
3/6/2010	5	21	22
3/13/2010	6	22	20
3/20/2010	4	24	15
3/27/2010	5	16	17
4/3/2010	4	16	20
4/10/2010	6	16	22
4/17/2010	7	16	21
4/24/2010	6	22	27
5/1/2010	9	19	20
5/8/2010	7	25	29
5/15/2010	7	25	26
5/22/2010	8	45	19
5/29/2010	12	37	25
6/5/2010	16	29	29
6/12/2010	17	28	31
6/19/2012	15	47	24
6/26/2010	***100***	47	62
7/3/2010	54	37	45
7/10/2010	35	**64**	45
7/17/2010	37	46	38
7/24/2010	**63**	37	51
7/31/2010	46	51	44
8/7/2010	42	41	34
8/14/2010	35	47	42
8/21/2010	37	***100***	45
8/28/2010	41	44	46
9/4/2010	37	37	44
9/11/2010	41	43	**61**
9/18/2010	**63**	**65**	**74**
9/25/2010	**59**	**67**	57
10/2/2010	39	**66**	53
10/9/2010	36	**61**	51
10/16/2010	37	58	56
10/23/2010	**46**	**68**	**70**
10/30/2010	39	57	**81**
11/6/2010	36	42	**100**
11/13/2010	30	53	**90**
11/20/2010	31	**74**	**70**
11/27/2010	21	**81**	**63**
12/4/2010	23	58	**67**
12/11/2010	21	43	41
12/18/2010	22	43	43
12/25/2010	16	36	33
1/1/2011	19	37	22

Descriptive Analysis of Twitter

There were forty-one Tweets identified as relevant to pertussis; each Tweet was further classified using a modified Vieweg approach (47), resulting in a small number of potential Tweets for further analysis (Table 6.8).

Table 6.8. *Count of Tweets per state that fall into modified Vieweg categories (47)*

Category	***California***	***Michigan***	***Ohio***
Relevant	10	1	1
Prevention	10	0	0
Article	4	0	1
Not Relevant	5	1	0
Duplicate or RT	7	1	0

The twelve Tweets deemed relevant were then reviewed in further detail to assess credibility following Gupta's methodology that classified them as definitely credible, seems credible, definitely incredible, or credibility unclear (46). Of the ten California Tweets deemed relevant, seven seem credible while three are unclear (Table 6.9). For Michigan and Ohio, there was only one Tweet for each state that was deemed relevant per state on June 8th and April 16^{th,} respectively. The credibility for both could not be determined.

Table 6.9. *Relevant Tweets for California with associated credibility rating*

Tweet Text	***Date***	***Credibility***
woke up dry whoop coughing up a lung and sounding like barry white...feel like shit but at the same time I feel fine. WTF?!? FML ugh ugh!	2/2/10	Somewhat
i had zero jobs for 4 months and now i have 2 ? apple just hired me ! woop woop *cough*! you all just caught whooping cough !	5/5/10	Unclear
@lizzhuerta Is it a "whoop" then a cough, or a cough then the "whoop"? Concerned people want to know...	7/15/10	Somewhat
Who ever heard of a non-contagious whooping cough??? "Whoop there it is"	7/31/10	Unclear
Every year everyone comes back from the gathering sick. Diagnosis? The whoop-whooping cough.	8/21/10	Somewhat
Whooping Cough - not just for Oregon Trail. Next I'll get dysentery or ford my wagon across a river. *whoop*	8/30/10	Somewhat
Rejected headline for weekend piece on whooping cough epidemic in Marin: "Whoop, here it is"	9/4/10	Somewhat
DS not breathing well tonight. Cough has an absolute "whoop" at the end. I think tomorrow am we go get tested for Pertussis…	9/18/10	Somewhat
I am not usually a kid person so I would like to know who gave me the whoop.	9/21/10	Somewhat
All my Juggalo homie's let me get a Whoop Whoop'ing Cough cuz we so sick. All my Gigolo homies well actually I don't want anythin' u got!	10/4/10	Unclear

For the five relevant Tweets containing links, the link was reviewed to see if that content needed to be included in this research. Two of the links were dead (leading to nonexistent webpages), and one link led to a Facebook posting by a local doctor in Ohio. Of the two remaining news articles referenced within the Tweets, one provided specific case counts for San Diego County, but no state counts. The other article did provide a state pertussis count [910] on June 24th. This article would be considered novel, but was not included in this research because this approach to gathering news stories was not part of the methodology.

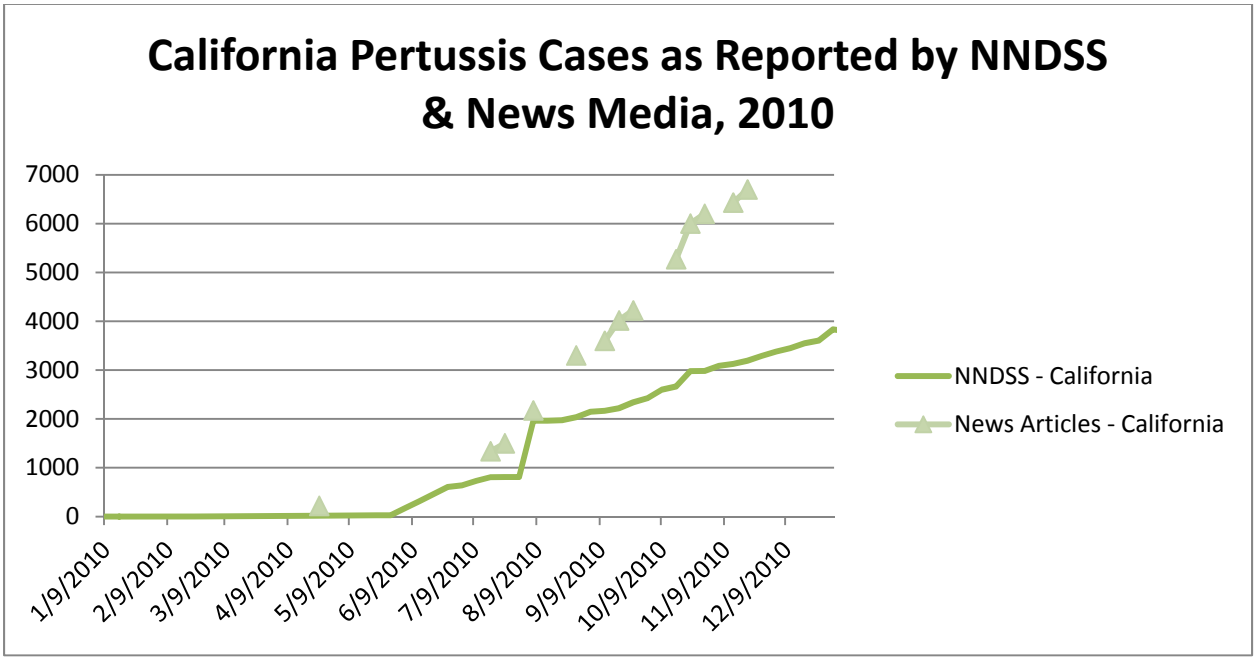
Comparative Analysis of NNDSS and News Articles

Timeliness

For comparison of NNDSS to news articles, the cumulative case count was used to create a pair of epidemic curves. A cumulative view must be used for news since articles are reporting the total number of cases rather than providing weekly snapshots of incidence.

When creating the comparison data set, news articles were aligned as closely as possible to NNDSS weeks with a news date preceding NNDSS week dates if necessary. If two news reports fell within one MMWR week, the larger number was used for news reporting for that week. This comparison was done for California and a graphic interpretation can be seen in Figure 6.4. This figure shows an earlier reporting of cases by the news media compared to official report dates confirming the role of news media as an early indicator of an emerging event. By the end of the year, the news articles indicated a case count nearly double that of official reports (NNDSS) which may indicate further reporting delays for official data, or could indicate misinformation within the news articles.

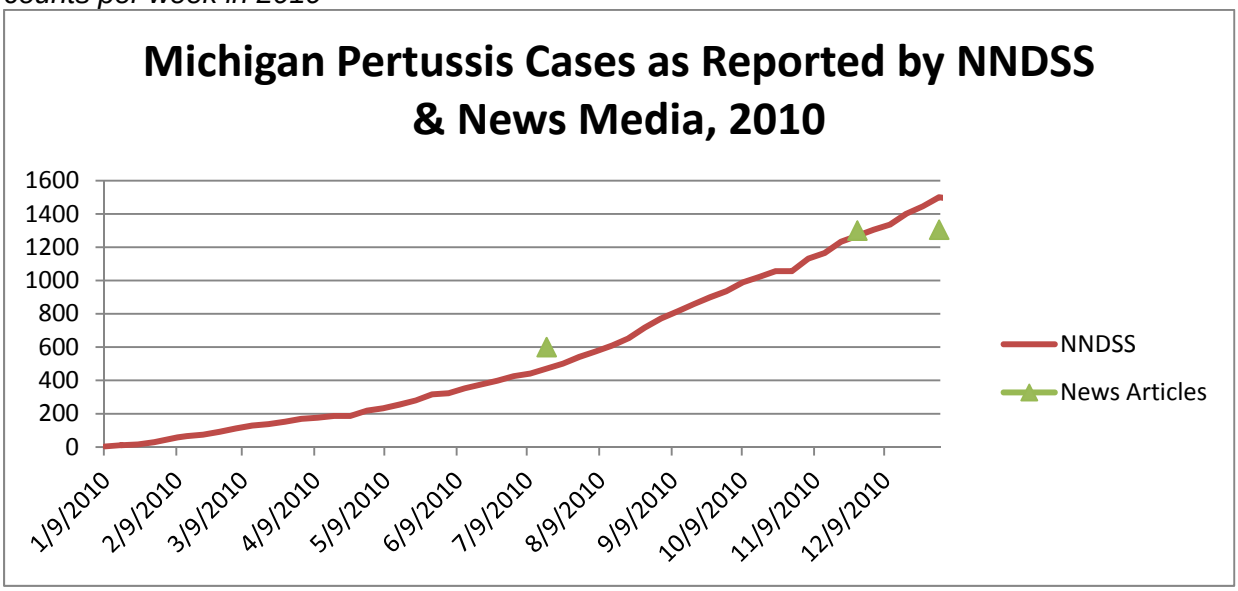
Figure 6.5. *Comparison of NNDSS epidemic curve for California and news article case counts per week in 2010*



The same comparison was done for Michigan, which showed different results and fewer news stories (Figure 6.5). For Michigan, news sources had an earlier report of the pertussis case counts in the summer during the week of July 17. The news reported 600 cases for that week, while NNDSS data indicated only 471 cases. Michigan would not reach

600 cases until one month later (the week of August 14). Later in the year, during the week of November 27, the news reports were close to the actual case counts [NNDSS reported 1,269 cases and the news reported 1,300]. During the last week of the year, the news sources underreported actual cases from NNDSS, which indicates that the news does not consistently have value throughout all phases of a disease outbreak.

Figure 6.6. *Comparison of NNDSS epidemic curve for Michigan and news article case counts per week in 2010*



For Ohio, there was only one identified news story identified that had an actual case count in it. That article, published December 2, reported 1,546 cases. The actual case count from NNDSS for that week (ending December 4) was 1,626. The conclusion we must draw is that, for Ohio, the news did not provide advanced notice of emerging pertussis cases.

For further comparison, the weekly case counts from NNDSS were compared to the news article reports (Table 6.10). "Negative timeliness" would be when fluctuations in the article case counts preceded fluctuations in the NNDSS, whereas positive results mean NNDSS data precedes search frequency (102). In California and Michigan, NNDSS weekly reporting of changes in rates preceded news article prevalence by weeks to months. In Ohio, NNDSS reported peaks of the pertussis outbreak at the same time as the news. In

addition, since the news reports did not occur each week in Ohio, there were fewer weeks from which to identify the peak making this comparison different from the comparison to Google.

Table 6.10. *Comparison of peak weeks between NNDSS and news articles*

	NNDSS Peak Week	News Peak Week	Difference (in weeks)	More Timely Source
California	31	46	+15	NNDSS
Michigan	44	52	+8	NNDSS
Ohio	48	48	0	Neither

Sensitivity and Specificity

The next analysis was to assess the sensitivity and specificity of news articles against NNDSS reporting, with the intent to assess the accuracy of the news reporting. The sensitivity value indicates the ability of news articles to identify true epidemic weeks (based on NNDSS), while the specificity value indicates the ability of news articles to identify true non-epidemic weeks (Table 6.11). For all three states, there were no "true positives" identified (no weeks where both NNDSS and news articles showed an increase), which resulted in both the sensitivity and positive predictive value being zero. This result means the news had no ability to identify true epidemic weeks. This predictive value may be lower than expected because for these calculations because the reporting weeks were aligned based on calendar dates, which rejects the hypothesis that news precedes NNDSS. With the positive predictive value (PPV) for all three states being zero, this shows an indication of an epidemic week from the news, but may or may not be indicative of an actual increased case count week in NNDSS.

Specificity calculations for all three states are high, indicating that case counts reported in the news are an accurate barometer for determining non-epidemic weeks. For all three states, the negative predictive value (NPV) is high, indicating that if news case counts indicate the week is not an epidemic week, it is highly likely (over 84%) that this

report is accurate. To assess accuracy, F-1 scores were calculated to evaluate the performance of news reports for identifying epidemic weeks in alignment with the NNDSS weekly case counts. For each state, the F-1 score was zero since precision and recall were both zero. This result means that the news had no ability to detect accurately pertussis outbreaks in California, Michigan, or Ohio.

Table 6.11. *Sensitivity, specificity, and predictive values of news articles to accurately identify NNDSS spikes*

	Sensitivity	**Specificity**	**Positive Pred. Value (PPV)**	**Negative Pred. Value (NPV)**	**Accuracy (F-1 Score)**
California	0%	94%	0%	96%	0
Michigan	0%	100%	0%	87%	0
Ohio	0%	98%	0%	82%	0

Correlation

Pearson's correlation coefficient was calculated to measure the strength of the relationship between the NNDSS data and the article case count (103). Pearson's attempts to draw a line of association through the data of two variables, with the r value indicating how far each variable is from the best fit; a perfect direct association exists if the value for the correlation coefficient is +1, while a perfect indirect association exists if the value is -1 (103). Table 6.12 provide both the correlation coefficient and the p-value of the null hypothesis that there is no correlation between the two sources (value is zero). The results indicate there is no correlation between NNDSS and news article reports in California or Michigan. Since there was only one value for news reports in Ohio, it was not possible to calculate correlation.

Table 6.12. *Correlation coefficients from comparing NNDSS data to news article case counts*

	Correlation Coefficient	Strength of Association	p-value
California	-.15	Small	.62
Michigan	.69	Large	.51
Ohio	-	-	-

How Accurate Were the News Media?

While it is important to look at the individual factors that reflect accuracy independently, in order to provide an overarching view of findings, the researcher developed Table 6-13 to summarize five tests of accuracy for news media for each state.

Table 6.13. *Summary of factors comparing NNDSS to news articles, (+) indicates news articles outperformed official reporting, (-) indicates NNDSS outperformed news articles, (o) indicates both Google and news articles performed equally.*

	California	Michigan	Ohio
Timeliness	-	-	o
Sensitivity	-	-	-
Specificity	+	+	+
Accuracy	-	-	-
Correlation between NNDSS and news articles?	No	No	No

Comparative Analysis of NNDSS and Google Search

Timeliness

The first step in comparative analysis was to assess each of the novel data sources against the official (gold standard) source to assess timeliness. A temporal graphic incorporating both the NNDSS data and the Google search trends data was developed for each of the three states (Figures 6.6 to 6.8) and a visual comparison of trends was completed. Because Google's search frequency is a value for a given week, in order to complete this comparison using similar units, the NNDSS weekly case count is the data source. For California and Michigan, there is a distinct early peak in Google search trends

as compared to the official reporting of pertussis, this temporal association is not as obvious in Ohio.

Figure 6.7. *Comparison between NNDSS weekly case count (blue) and Google Insight search frequency (gray) in California per week.*

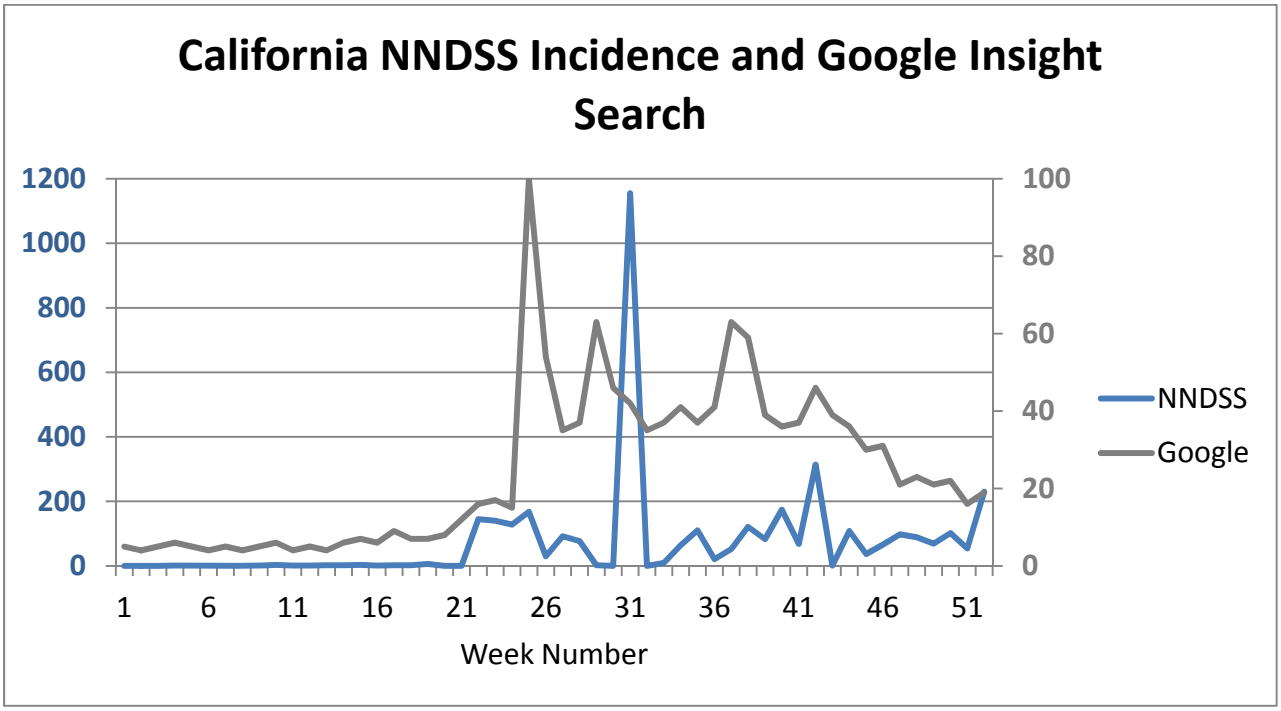


Figure 6.8. *Comparison between NNDSS weekly case count (red) and Google Insight search frequency (gray) in Michigan per week.*

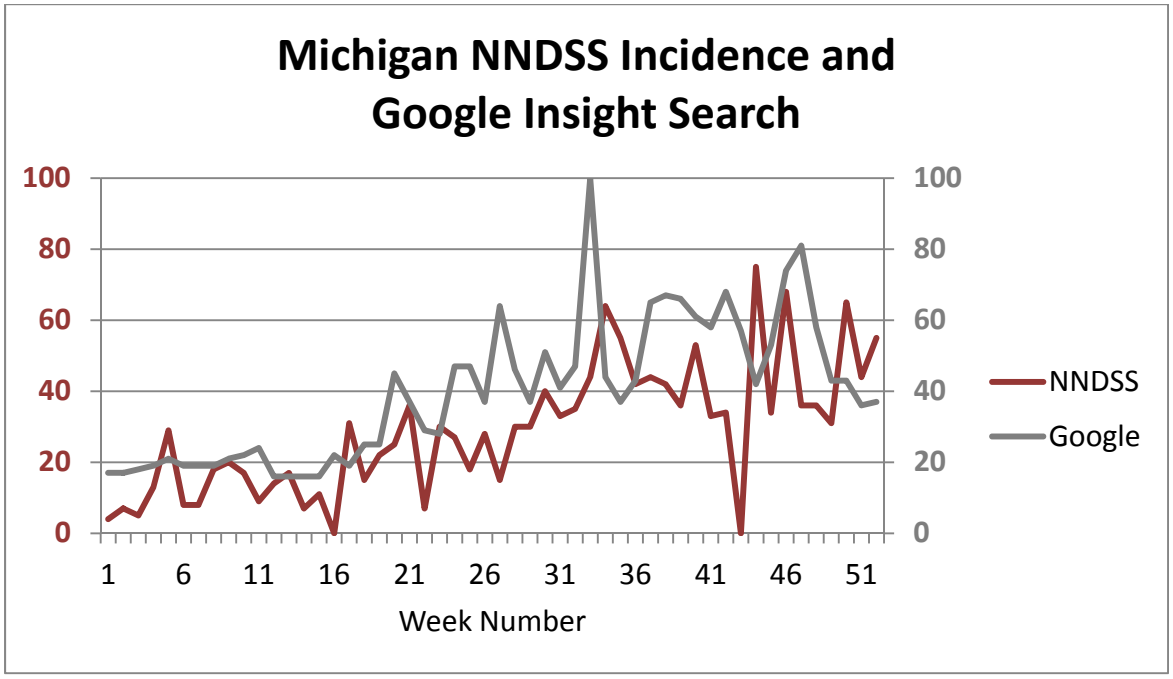
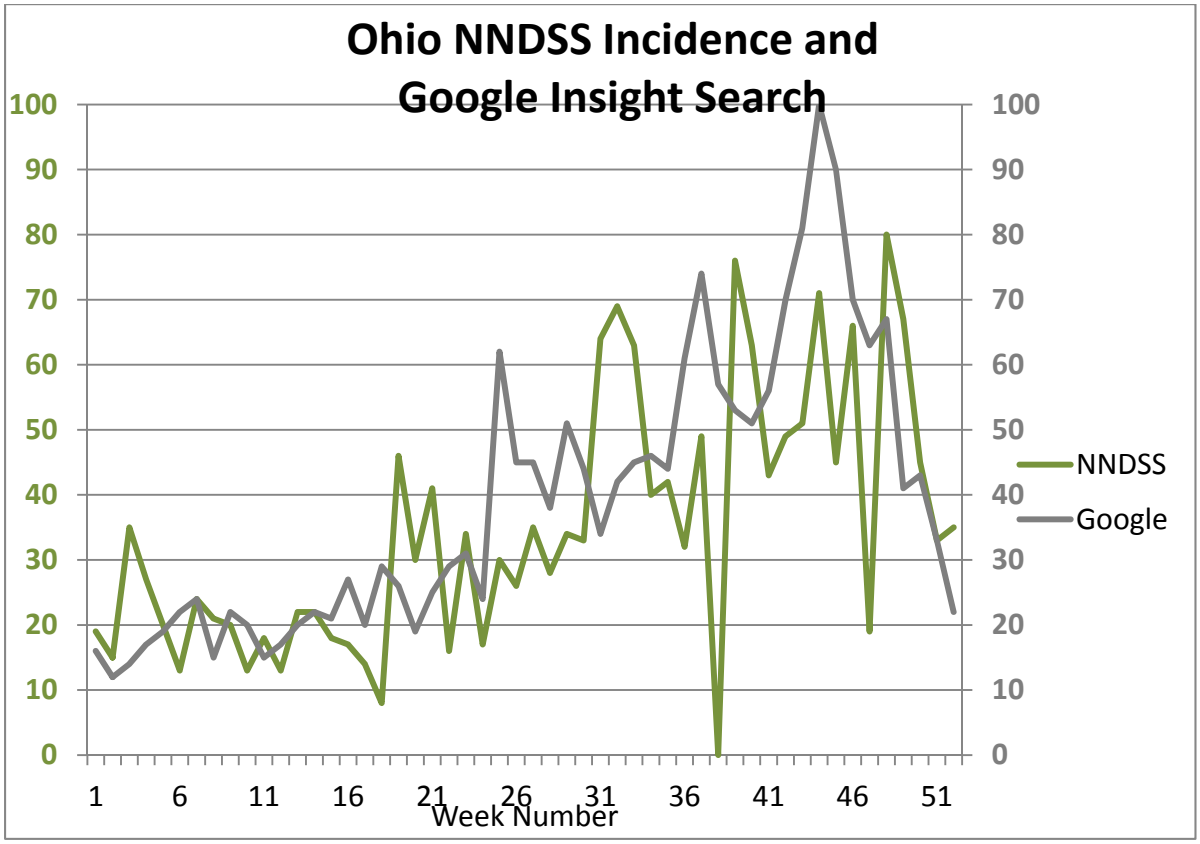


Figure 6.9. *Comparison between NNDSS weekly case count (green) and Google Insight search frequency (gray) in Ohio per week.*



For further comparison, the maximum weekly case counts from NNDSS were compared to the peak news-reported case count (Table 6.14). Negative timeliness indicates that fluctuations in internet search preceded fluctuations in the NNDSS, whereas positive results mean NNDSS data precedes internet search (102). In each state, Google search frequency preceded NNDSS by weeks to months.

Table 6.14. *Comparison of peak weeks between NNDSS and Google*

	NNDSS Peak Week	Google Search Peak Week	Difference (in weeks)	More Timely Source
California	31	25	-6	Google
Michigan	44	33	-11	Google
Ohio	*48*	*44*	*-4*	*Google*

Additionally, it was important to look at the first (earliest) week where the datum was one standard deviation or greater than the annual mean. This week would serve as the earliest epidemic intelligence that an epidemic may occur (although it is not a definitive sign). In California, the peak search frequency week was the earliest week that exceeded one standard deviation (and that week exceeded three standard deviations); additionally, there was a week (Week 29) that was above one standard deviation above the mean two weeks before the NNDSS peak (Figure 6.9). In Michigan, the first time the search frequency exceeded one standard deviation above the mean occurred six weeks prior to the search frequency peak, and seventeen weeks prior to the NNDSS peak at Week 44 (Figure 6.10). This peak provided almost four months advanced notice of a variation from the normal weekly counts. In Ohio, the first time the search frequency exceeded one standard deviation occurred eight weeks prior to the search frequency peak and twelve weeks prior to the NNDSS peak (Figure 6.11). This search provided almost three months advanced notice of a variation from the normal weekly counts.

Figure 6.10. *NNDSS Weekly Case Count for California with vertical lines indicating peak weeks from Google search frequency data in California (brown indicates absolute search peak, yellow indicates search frequency was one standard deviation above the mean)*

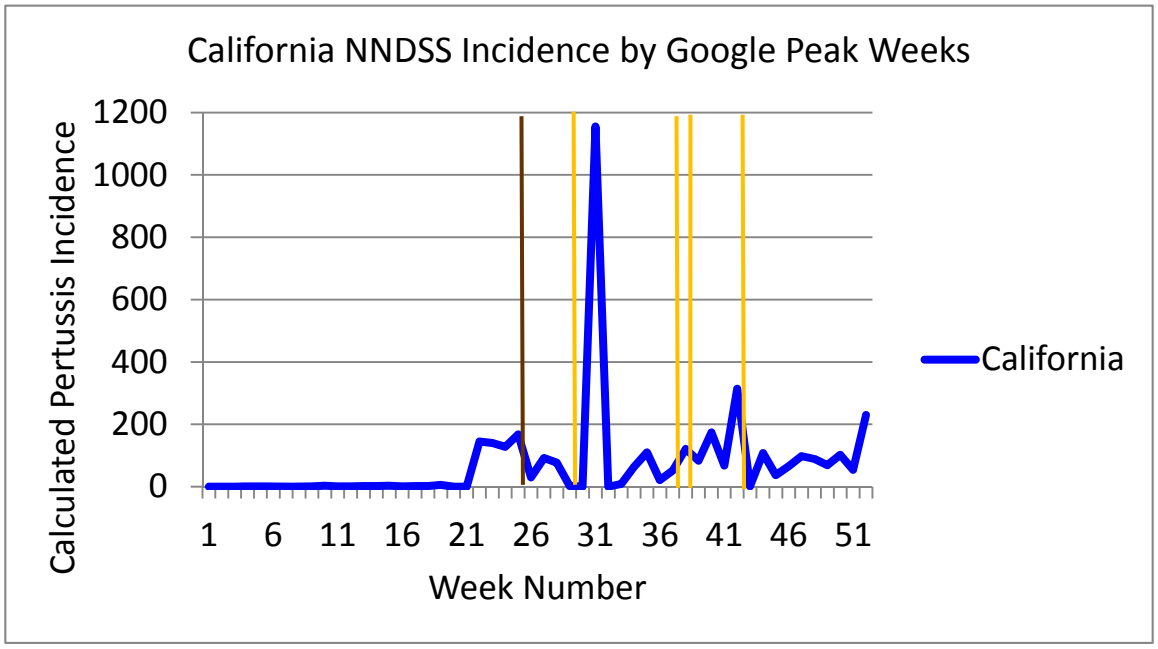


Figure 6.11. *NNDSS Weekly Case Count for Michigan with vertical lines indicating peak weeks from Google search frequency data in California (brown indicates absolute search peak, yellow indicates search frequency was one standard deviation above mean)*

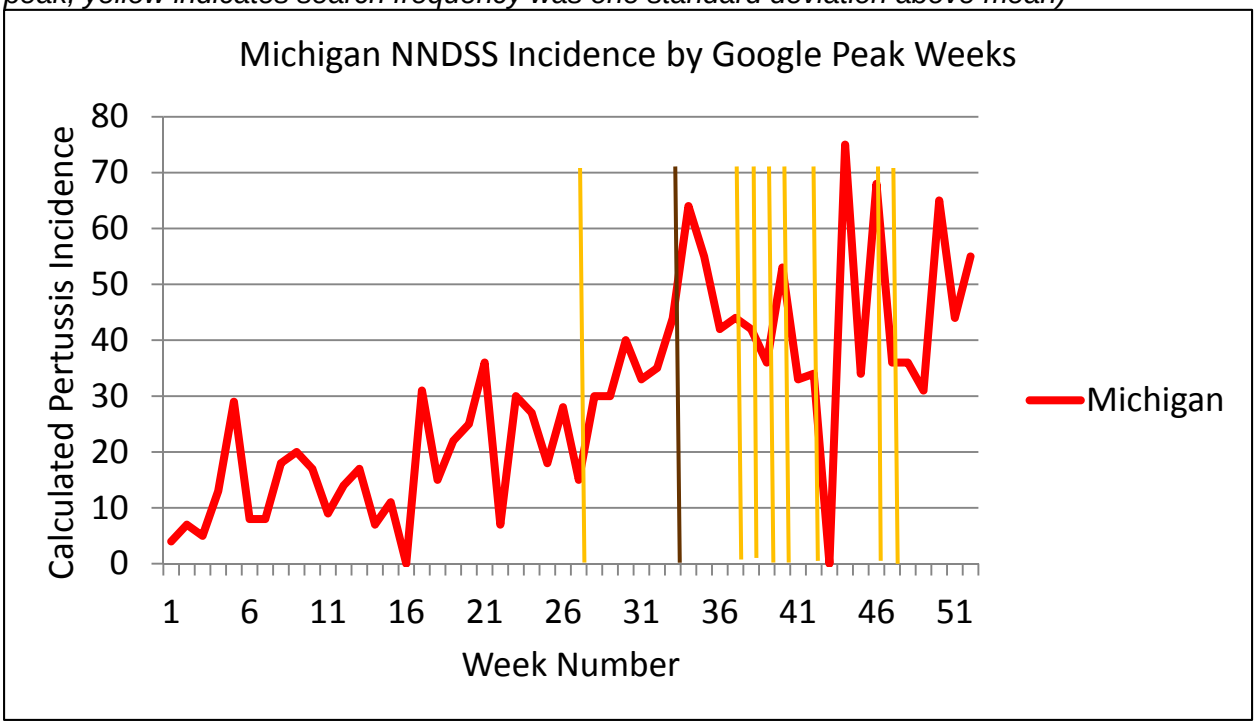
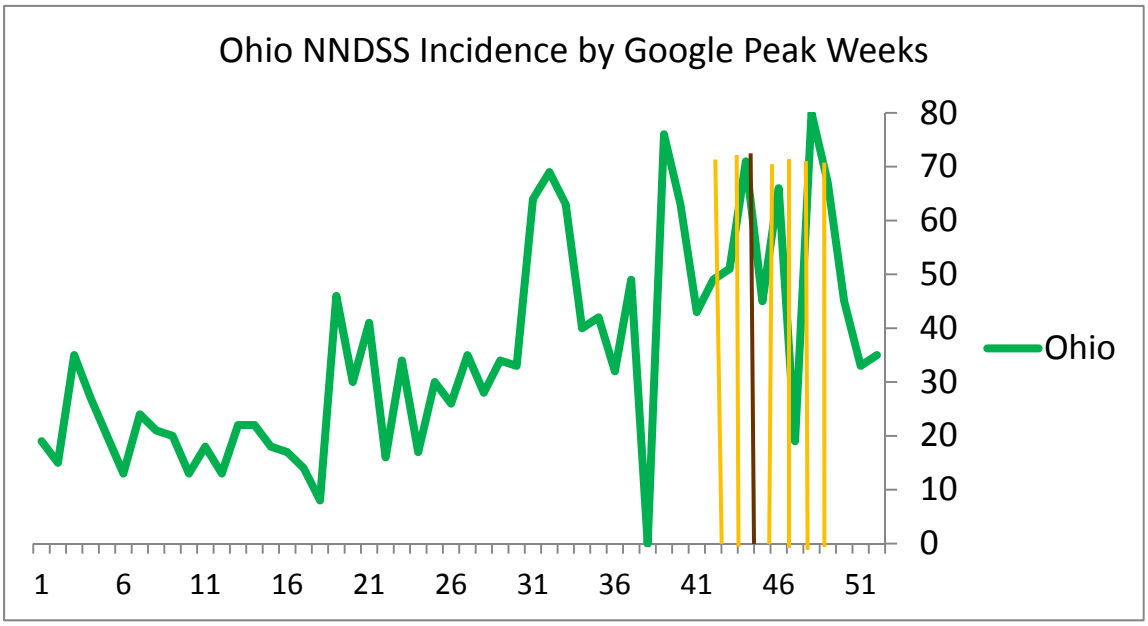


Figure 6.12. *NNDSS Weekly Case Count for Ohio with vertical lines indicating peak weeks from Google search frequency data in Michigan (brown indicates absolute search peak, yellow indicates search frequency was one standard deviation above mean)*



Sensitivity and Specificity

The next analysis was to assess the sensitivity and specificity of Google search frequency against NNDSS reporting; this analysis is essentially an assessment of the accuracy of the Google data set. The sensitivity value indicates the ability of Google search frequency to identify true epidemic weeks (based on NNDSS), while the specificity value indicates the ability of Google search frequency to identify true non-epidemic weeks (Table 6.15). As indicated, there is variation in both measures between each California, Michigan, and Ohio, with Google data in all three states having a high likelihood of identifying non-epidemic weeks. In California, there is a 50% likelihood that an increase in Google search volume is indicative of an increase in actual cases (i.e. 50% chance that Google identifies true positives). This likelihood may be lower than expected because these calculations were assessing the likelihood that the Google values and the NNDSS values would match (increases on the same weeks), which is against the hypothesis that Google

reporting precedes NNDSS. Sensitivity calculations for both Michigan and Ohio are relatively poor, indicating that Google indicates many false positives.

Both positive and negative predictive values were separately calculated for California, Michigan, and Ohio (Table 6.15). For all three states, the negative predictive value (NPV) is high, indicating that if Google search frequency indicates the week is not an epidemic week, it is highly likely (over 85%) that this indication is accurate. The positive predictive value (PPV) for all three states is lower, so an indication of an epidemic week from Google may or may not be indicative of an actual increased case count week in NNDSS. To assess accuracy, the F-1 scores were calculated to evaluate the performance of Google search frequency for identifying epidemic weeks in alignment with the NNDSS weekly case counts. These scores indicated low accuracy of Google search frequency as a measure.

Table 6.15. *Sensitivity, specificity, and predictive values for Google Search Frequency to accurately identify NNDSS spikes*

	Sensitivity	**Specificity**	**Positive Pred. Value (PPV)**	**Negative Pred. Value (NPV)**	**Accuracy (F-1 Score)**
California	50%	92%	20%	98%	.29
Michigan	29%	84%	22%	88%	.25
Ohio	22%	93%	40%	85%	.29

Correlation

Pearson's correlation coefficient was calculated to measure the strength of the relationship between the NNDSS data and the Google search frequency (103). Pearson's attempts to draw a line of association through the data of two variables, with the r value indicating how far each variable is from the best fit; a perfect direct association exists if the value for the correlation coefficient is +1, while a perfect indirect association exists if the value is -1 (103). The results in Table 6.16 provide both the correlation coefficient and the p-value of the null hypothesis that there is no correlation between the two sources (value is

zero). The results indicate a moderate positive correlation in both Michigan and Ohio which is statistically significant (p<.0001).

Table 6.16. *Correlation coefficients from comparing NNDSS data to Google search frequency*

	Correlation Coefficient	Strength of Association	p-value
California	.29	Small	.04
Michigan	.54	Large	<.0001
Ohio	.64	Large	<.0001

Comparison Characteristics for Search Terms

While it is important to look at the factors independently in order to provide an overarching view of findings, the researcher developed Table 6.17 to determine if there is variation between characteristics for internet search patterns within different states.

Table 6.17. *Summary of factors comparing NNDSS to Google Search, (+) indicates Google outperformed official reporting, (-) indicates NNDSS outperformed Google, (o) indicates both Google and NNDSS performed equally.*

	California	Michigan	Ohio
Timeliness	+	+	+
Sensitivity	-	-	-
Specificity	+	+	+
Accuracy	-	-	-
Correlation between NNDSS & Google?	No	Yes	Yes

Comparative Analysis of NNDSS and Twitter

Timeliness

In California, the first Tweet of the year regarding pertussis came during Week 5, over six months prior to the peak week (Table 6.18). There was also one week with two Tweets (Week 35), while each of the other weeks had only one Tweet (or zero); if multiple Tweets in a given week was the required indicator, than NNDSS would precede Twitter. For

both Michigan and Ohio, there was only one relevant pertussis Tweet per state with unclear credibility, and the dates for these Tweets were June 8 and April 16, respectively (Table 6.18). For Michigan, the Tweet came during Week 22, five months prior to the peak and three months prior to the first week in the NNDSS that was one standard deviation above the average (Week 34). In Ohio, the single Tweet occurred during Week 14, a full nine months prior to the peak and four months prior to the first week that was one standard deviation above the average (Week 31) in NNDSS.

Table 6.18. *Comparison of peak weeks between NNDSS and Twitter*

	NNDSS Peak Week	Twitter Peak Week	Difference (in weeks)	More Timely Source
California	31	5	-26	Twitter
Michigan	44	22	-22	Twitter
Ohio	48	34	-14	Twitter

Sensitivity and Specificity

Both positive and negative predictive values were separately calculated for California, Michigan, and Ohio (Table 6.19), to include calculating California values in two circumstances: one where seemingly credible Tweets only were included and one set of calculations for if both credible and unclear Tweets were included. For all three states, the negative predictive value (NPV) is high, indicating that if there are no Tweets in a given week then the week is not an epidemic week, and it is highly likely (over 80%) that this indication is accurate. The positive predictive value (PPV) for all three states is zero indicating that if there is a Tweet on a given week it is not indicative of an actual increased case count within NNDSS. Because there were no true positives (weeks where NNDSS data spiked and there was a Tweet), the F-1 score for accuracy was zero.

Table 6.19. *Sensitivity, specificity, and predictive values for Twitter to accurately identify NNDSS spikes*

	Sensitivity	Specificity	Positive Pred. Value (PPV)	Negative Pred. Value (NPV)	Accuracy (F-1 Score)
California-Credible	0%	88%	0%	96%	0
California-All	0%	82%	0%	95%	0
Michigan	0%	98%	0%	86%	0
Ohio	0%	98%	0%	82%	0

Correlation

Because there was only a single Tweet on any week for Michigan and Ohio and two Tweets in a single week for California, calculating correlation was not done for any of the states.

Comparison Characteristics for Twitter

In a summary for Twitter like those for news media and internet search, the researcher developed Table 6.20 to determine if there is variation between characteristics for social media within different states.

Table 6.20. *Summary of factors comparing NNDSS to Twitter, (+) indicates Twitter outperformed official reporting, (-) indicates NNDSS outperformed Twitter, (o) indicates both Twitter and NNDSS performed equally.*

	California	Michigan	Ohio
Timeliness	-	+	+
Sensitivity	-	-	-
Specificity	+	+	+
Accuracy	-	-	-
Correlation between NNDSS & Twitter?	Inconclusive	Inconclusive	Inconclusive

Combined Analysis (NNDSS and All Infodemiology Sources)

Timeliness

The primary way to look at timeliness of all the infodemiology sources is to look at the first indication (or signal) from any of the sources on the same graph along with the epidemiology curve (Figures 6.12 – 6.14), as first done by Keller (104) . This graph was

done per state, to see more easily what emerged in each of the three environments. In California, each of the three indicators from infodemiology data arose in advance of the epidemic taking off within the state. In Michigan, all three indicators also occur prior to the epidemic peak; however, the indicators are closer to the epidemic uptake than in California. Lastly, in Ohio all of the indicators arose while the epidemic was underway, indicating that, in that state, the utility of infodemiology sources in providing an early signal of an emerging outbreak was not great.

Figure 6.13. *Pertussis cases in California (based on NNDSS data) with vertical lines indicating the first signal from social media (purple), second from news media (blue), and third from internet search patterns (green)*

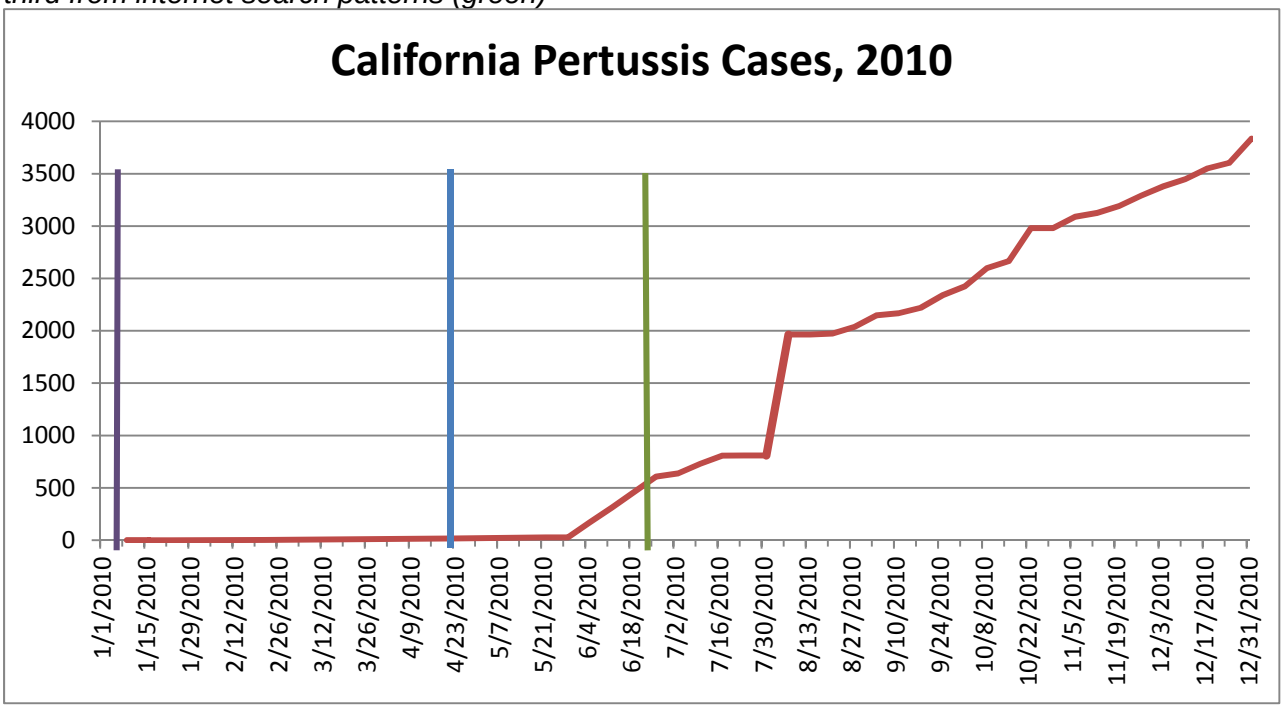


Figure 6.14. *Pertussis cases for Michigan (from NNDSS data) with vertical lines indicating the first signal from social media (purple), the second from internet search patterns (green), and the third from news media (blue)*

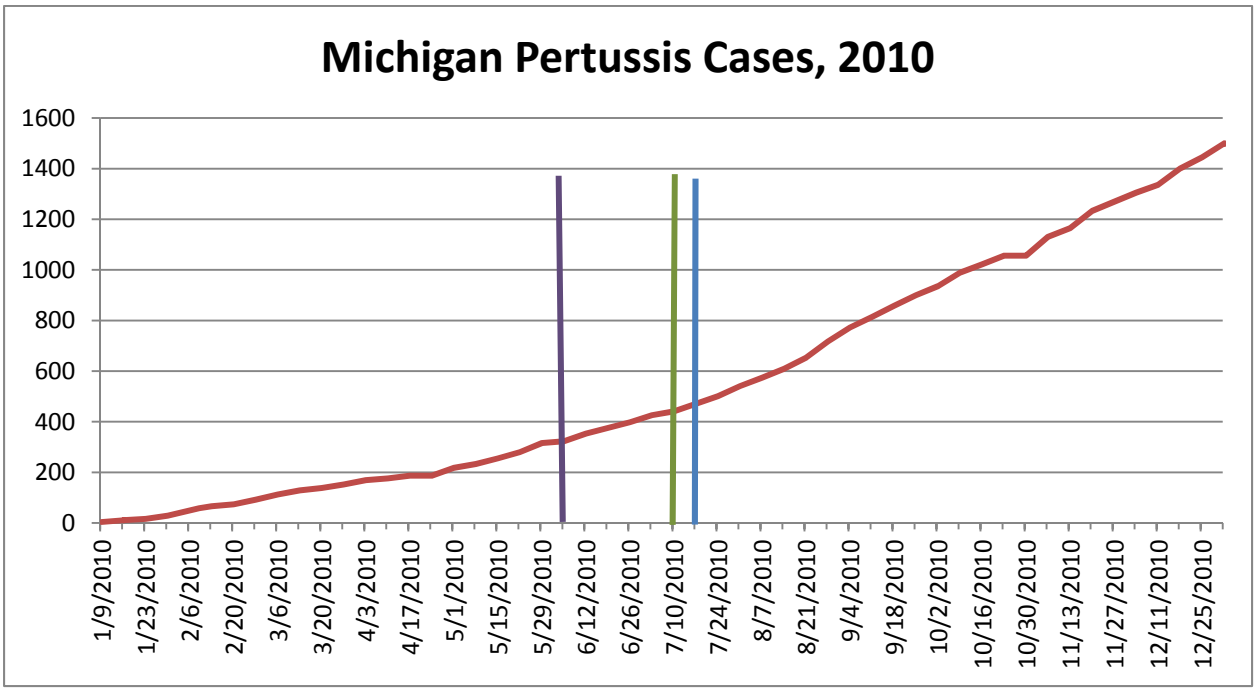
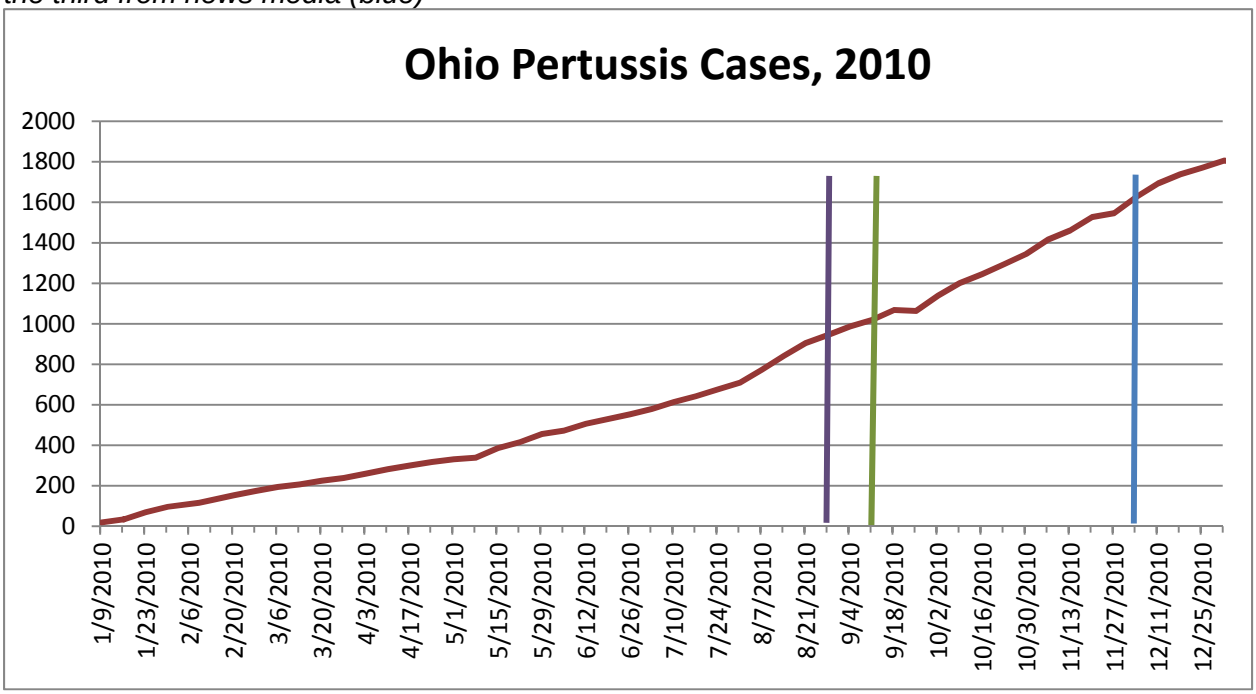


Figure 6.15. *Pertussis cases for Ohio (from NNDSS data) with vertical lines indicating the first signal from social media (purple), the second from internet search patterns (green), and the third from news media (blue)*



Despite the extremely low frequency of Tweets regarding pertussis, single Twitter posts arose weeks ahead of both news and internet search indicators in all three states. Internet search patterns and news media followed behind in that order in Michigan and Ohio, with the two reversed in California. From this limited research, it is not possible to determine if the infodemiology source that first reported information would impact the way, likelihood, or timeliness of other sources due to the scarcity of the initial source (Twitter).

Sensitivity and Specificity

As a way to assess sensitivity and specificity of any of the infodemiology sources as compared to NNDSS, we calculated values (Table 6.21). Only one of the sources (news media, internet search patterns, or social media) had to indicate a spike to be considered as an event in this calculation. This variable was done to understand that if a decision maker had access to all of these sources (and considered each to be equal), how would this modify the ability to detect prior to NNDSS. In California, only Google data had an actual true positive, so the values are better than for either news or social media alone. In Michigan, the accuracy score from using all three sources is higher than the score of any single source, indicating that a combined look at infodemiology sources provides the best insight into events ongoing as compared to news, internet search, or social media alone. In Ohio, the accuracy score for combined infodemiology sources is better than for news media or social media alone but not as high as internet search patterns.

Table 6.21. *Sensitivity, specificity, and predictive values for news media, internet search, or Twitter to accurately identify NNDSS spikes*

	Sensitivity	**Specificity**	**Positive Pred. Value (PPV)**	**Negative Pred. Value (NPV)**	**Accuracy (F-1 Score)**
California	50%	72%	7%	97%	0.12
Michigan	40%	82%	25%	88%	0.31
Ohio	22%	88%	29%	84%	0.25

Analysis of Survey Results

A survey (see Appendix C) was distributed to 521 members of the International Society of Disease Surveillance (ISDS), and there were 71 complete responses to the survey, resulting in a response rate of 13.63%. Only completed survey results are included in the analysis. The respondents worked in a variety of settings, with 50% working within a state or city/county health department. Amongst the respondents, most (66%, n=46) were between the age of thirty and forty-nine. The first question in the survey asked respondents to identify which popular internet tools they had used (Table 6.22).

Table 6.22. ***Summary of responses to question on internet tool usage***

Which of the following have you ever used (select all that apply):	***n***	***%***
Twitter	37	52
Facebook	55	77
Search Engines	69	97
Email News Alerts (like ProMED)	64	90

When looking at usage over the past month, the most used technology (used more than ten times) was the search engines with 96% (n=66), followed by both Facebook (51%, n=28) and email news alerts (56%, n=36). Alternatively, for Twitter, the most common usage response was one to two times over the past month (35%, n=13).

A majority of the respondents agreed with the statement "Community members can provide valuable information about disease outbreaks in their community," with 83% (n=59) agreeing. The variation was with the level of agreement which was split closely, with 38% (n=27) somewhat agreeing and 45% (n=32) strongly agreeing. Confirmatory data was considered necessary by 89% (n=63) of respondents, but the variability comes when differentiating between usually necessary (63%, n=45) or always necessary for (25%, n=18). Amongst respondents, 80% (n=57) indicated they were always looking for new types of data and information sources to inform public health response actions. There was more variation

in the question focused on the possibility of getting too much information during an event. Responses were mixed from 6% strongly agreeing, to 38% agreeing, 25% disagreeing, 15% strongly disagreeing, and 15% unsure.

The key questions for this study were about the utility of news media, social media, and internet search patterns to inform the three stages of situational awareness discussed earlier. The results (Table 6.23) indicate that the utility of these novel sources varies based on the stage of situational awareness (identifying early indications or signals).

For most survey respondents, infodemiology content had the most value in the first stage of situational awareness for identifying early indications of disease outbreaks. News media and internet searches were moderately to highly valuable for 70% of respondents, while social media was moderately to highly valuable to 60% of respondents. For both strengthening comprehension of an outbreak and informing future predictions, beliefs were split regarding the level of potential value (if any) that exists.

Table 6.23. *Summary of findings for survey questions regarding value of social, news, and internet search for emerging disease outbreaks, red font indicates highest value per row.*

	Highly Valuable	**Moderately Valuable**	**Minimally Valuable**	**Not Valuable**	**Unsure**
How valuable in identifying early indications or signals of emerging disease outbreaks is information from					
Social Media	13 (18%)	**30 (42%)**	16 (23%)	6 (8%)	6 (8%)
News Media	20 (28%)	**30 (42%)**	14 (20%)	3 (4%)	4 (6%)
Internet Search Patterns	20 (28%)	**29 (41%)**	13 (18%)	5 (7%)	4 (6%)
How valuable in strengthening your comprehension of emerging disease outbreaks is information from					
Social Media	8 (11%)	19 (27%)	**24 (34%)**	12 (17%)	8 (11%)
News Media	15 (21%)	**28 (39%)**	18 (25%)	7 (10%)	3 (4%)
Internet Search Patterns	12 (17%)	**26 (37%)**	18 (25%)	9 (13%)	5 (7%)
How valuable for informing future predictions of emerging disease outbreaks is information from					
Social Media	7 (10%)	**22 (31%)**	16 (23%)	11 (15%)	15 (21%)
News Media	11 (16%)	**23 (32%)**	11 (16%)	15 (21%)	10 (14%)
Internet Search Patterns	16 (23%)	**19 (27%)**	15 (21%)	9 (13%)	12 (17%)

When looking at infodemiology content to identify early signals of disease outbreaks, all three potential sources were seen as moderately valuable by about 40% of survey

participants. The distinction is between social media, seen as highly valuable by 18% and both news and internet search seen as highly valuable by 28%. This distinction was not uncommon in either of the other two phases of situational awareness (strengthening comprehension or future prediction). When assessing the value of social media for strengthening comprehension of an emerging outbreak, social media was most often seen as minimally valuable by participants. The area where there seemed to be the most uncertainty for the utility of infodemiology was in informing future predictions. This is also the phase of situational awareness where survey respondents had the opinion that these information sources were deemed not valuable, as compared to other phases where the content was deemed as at least minimally valuable.

There was a numeric value assigned to each response: 1=highly valuable, 2=moderately valuable, 3=minimally valuable; 4=not valuable. In Table 6.24, the researcher calculated the mean response given per question. All "Unsure" responses were eliminated from this calculation.

Table 6.24. *Comparison of mean value across types of media for three stages of situational awareness*

	Identifying Early Indicators	**Strengthening Comprehension**	**Informing Future Predictions**
Social Media	2.23	2.63	2.55
News Media	2.00	2.25	2.50
Internet Search Patterns	2.04	2.37	2.29

When looking across all phases of situational awareness, identifying early indicators of disease was the phase in which infodemiology sources were seen as most valuable in comparison to strengthening comprehension or informing prediction. The above results indicate that overall, news media were considered most useful as compared to any other

information source at a mean of 2.0. News media were seen as the most valuable for identifying early indicators and strengthening comprehension, while internet search patterns were most valuable for informing future predictions. With future predictions as the exception, news and internet search had very similar values and were seen to have similar utility in both Tables 6.23 and 6.24. Alternatively, social media was seen as less valuable for all phases of situational awareness, as compared to either news media or internet search patterns.

The survey also included an open-ended question focused on current knowledge gaps or information needs amongst survey respondents regarding novel data sources for situational awareness. Specifically, respondents were asked "*What questions need to be answered or research needs to be completed for you to have more complete insights about novel data sources*?" Most responses were around the need for additional evaluation or validation efforts focused on the strengths and limitations of the data with responses such as:

> *"They need to be assessed against sources on which we have relied up till now for decision-making, like reportable disease or syndromic surveillance data based on ED visits. We had to do this with SS* [syndromic surveillance] *data when it was new."*

Respondents wanted to understand how infodemiology performs under "*normal and outbreak conditions*," how it performs with low incidence diseases (which this research effort will help answer), and how it performs compared to syndromic surveillance.

Another key gap was in understanding how well infodemiology sources cover populations and what the variations are in coverage between the sources as well as between different locations to better understand "*Is the data source representative*?" Survey respondents also identified the need for better understanding of the potential applications of these data sources in regular practice. There was a specific request for efforts to attempt

using these sources in real-time rather than retrospective analyses, with one respondent indicating: "*Anyone can find anything when only looking back in time and data*." The need to understand the enhanced value this additional source adds compared to existing information sources at a time of limited resources was clearly articulated. Respondents also were hoping for additional insights on how (and when) to trust a source, based on some form of reliability assessment and some sort of clarity on the security of using such sources inquiring: "*Does it provide actionable information or is it just 'interesting?*"'

The final major topic discussed by many respondents was the need to understand how to differentiate between disease occurrence and interest in the disease. The request was to understand:

> "*How sources (such as news media) influence others (such as Facebook or Twitter). In other words, how can we remove rumor effects from actual signals? If diseases are discussed on news or social media outlets, does that result in increased search frequency or other terms?"*

And, if so, would it be possible to distinguish between the actual disease increasing in a community as compared to social contagion.

A second question was asked regarding what respondents need to know or understand before using a new information source. The responses identified many of the same gaps in understanding of the validity, reliability and credibility of the infodemiology sources. The biggest critical need was to understand how these sources performed in comparison to other sources, specifically "*What information or sources can be used to corrobate it?*" This was both looking at other sources available within that organization, as well as past outbreaks and the source's "*previous performance*".

Respondents also described a need to understand specifics of the data itself: how was it collected, by whom, when (timeliness), and who funded the collection of the data.

The other major category of need for information before using a novel source was specifics on the biases and limitations of the data, as well as a thorough understanding of what the data baseline values had been prior to a given event. One respondent was also focused on knowing the data utility and limitations at different phases:

> "*Sensitivity/specificity during each phase of the outbreak. While Twitter/Facebook may be appropriate to use at one phase (for example, let's say at the beginning), they may not be appropriate to use during other phases (like, indicating when the outbreak is over, or vice versa).*"

Lastly, the need for understanding how representative the data was of the overall population was mentioned by many responders, which aligned with broad understanding of the source and its related metadata.

CHAPTER 7. DISCUSSION

NNDSS Data

This section will discuss the limitations of each of the data sets as well as interpret the meaning of the trends as associations. The analysis of the NNDSS data overall allowed me to identify the temporal trends of pertussis infection reports and provide a baseline comparison for alternative infodemiology sources. Each state's patterns will be described and compared.

The California NNDSS data resulted in an epidemic curve that followed the "signature" outbreak pattern of a propagated epidemic that had a primary spike during Week 31, and a less significant secondary spike at Week 42. There were no other weeks of significant case counts or spikes within California, making the comparisons to infodemiology sources focus on a limited period. Michigan and Ohio both have multiple spikes that exceeded expected case counts (greater than one standard deviation above the mean), so it is not as evident where the primary (and related secondary or tertiary) spike was. In those states, the analytic focus needs to be on both the first spike that was significant and the spike that resulted from the highest raw case count (Weeks 34 and 44 for Michigan, Weeks 31 and 48 for Ohio).

With the values for cumulative counts not agreeing with previous weeks' sums plus current week counts, the researcher recalculated the past week's case count (see Chapter 5 Methods for Data Collection and Processing: Official Reporting). This calculation may or may not have resulted in case counts correctly associated with the appropriate week of

incidence, since it is unknown during what week the case actually occurred. This method associates new cumulative cases [not previously linked to a week] with the previous week, and this may have resulted in an NNDSS data set that pushed cases forward in time rather than accurately associating them with earlier weeks in the calendar year. Since the data from NNDSS are still provisional for 2010, the data may not be completely accurate since collection and aggregation are ongoing (92). Because of the contingent nature of the data, there is no way to determine which week was the accurate week for these case counts, or if the method the researcher had chosen was inaccurate.

One other potential implication of the cumulative case count not equaling the past cumulative total plus the current week is that situational awareness of public health officials may be compromised. During an ongoing event, the numbers (weekly case counts) would not have accurately equaled the cumulative total within a state. This cumulative total is critical in assessing impact of a disease on a community (or in this case, a state). The retrospective weekly calculated case counts were likely more accurate than the real-time weekly numbers, which emphasizes that the official source would not be accurate and should not be the only information source for decision makers. Perhaps, this gap is another critical reason for utilizing infodemiology content.

News Articles

Pertussis in California was an ongoing news story throughout the year with seventeen unique stories with statewide case counts; there were only a few articles in Michigan and Ohio. This figure does not take into account the proportion of pertussis stories against all stories about health threats or "breaking" news for each of these states, so it cannot be determined if pertussis was deemed newsworthy in comparison to other events over the course of the year. The news reporting for California showed a gradual upswing in total cases being reported. There was a single date where two newspapers reported two

different numbers (August 25, 2010). Possibly, due to the small number of reports in Michigan and Ohio, an upswing was not as apparent.

News stories compete for space and sensational events may get more coverage due to reader interest rather than newsworthiness (24). The "salience" of news is determined by unexpectedness, proximity, conflict, discrepancy, prominence, or celebrity status of involved people (105). The six obstacles to journalistic objectivity that may influence news content, resulting in bias, are values inherent in political events, deceptions by the newsmakers themselves, difficulty in achieving overall neutrality, impossibility of covering all sides and gathering all facts, rush to meet deadlines, and pressures of the continuous or 24/7 news cycle (106). The researcher did not interact with the writers or editorial boards of the papers indicated (and there is no written summary of such decisions to review), so it is not apparent from available information if any of these factors influenced reporting.

Of the articles incorporated in this research, most (fifteen of twenty-four) were from national news organizations who have internet-only presence or release their content over multiple formats. Of these articles, three were from the local affiliate of a national news organization. Seven articles were from local news organizations that had an internet-only presence. This result aligned with the continuing decrease in local news for all topics, and a more significant proportion of news coming from national organizations via an internet-only format. This result should be considered a positive from the perspective of accurate news, since local news has the potential to be less reliable, due to limited resources and training, or may be inadequately confirmed (9). Each geographic region of the country also has papers regarded as having a more liberal or conservative approach (75). Of the articles incorporated in this analysis, most (eighteen) were from news organizations that are not considered to have a clear political leaning that indicates that fear of, or attention to, disease may be apolitical.

A disease's mortality level affects the likelihood of a journalist writing about a disease, with the affected population generating more or less coverage (7). Because pertussis often impacts those considered vulnerable in our society (namely young children), it is expected that this disease is more likely to gain news coverage, but no comparison of pertussis news articles to other disease articles was done for this time period. It is also not apparent if the readership of the papers that had articles was the same as the population at risk, so the level of concern from news counts cannot be assumed to accurately reflect level of concern (72).

News Articles and NNDSS Data Comparison

Comparisons between news article case reporting and NNDSS weekly case reporting did not result in any clear advantage to using news articles as compared to NNDSS data in California, Michigan, or Ohio during 2010. There was also no correlation between the case counts reported by news and those reported officially.

In comparing news article case counts with NNDSS for California, the news was timelier; however, the news reported more cases than were reported officially. There were four news articles indicating a pertussis outbreak that occurred in advance of the primary spike (maximum weekly case count). Three of these reports included case count numbers that were above the actual weekly maximum that occurred in Week 31. The news included nine additional reports with increasing case counts between the primary and secondary spike. Conversely, it is possible that these (and other) findings could be coincidental, since there is no way to determine if a causality relationship exists between disease outbreaks and news reporting (7).

Based on these findings, the necessary step was to determine if the news would prove a sentinel indicator of an emerging epidemic by providing a single point of data. For instance, in 2009 California had 896 pertussis cases over the course of the year; in 2010, the state had reported this many cases by early August (4, 91). This statistic makes any

report indicating it may be an epidemic year prior to August extremely valuable in gaining situational awareness in advance of a confirmed epidemic (from a retrospective analysis). There was a single report in April 2010 indicating over 200 pertussis cases, and this report is the earliest identified signal to an emerging event. There were also two reports in July that provided indications of increased case counts (providing case counts that were above those reported for the total of the prior year).

In Michigan, the news case count was timelier than official reports in the summer of 2010, and the official report was timelier than the news in the winter. In Ohio, there was only a single news report with a pertussis case count included, and that report did not appear before official reports. In both Michigan and Ohio, indications of an emerging pertussis epidemic would likely have been detected by state-based surveillance systems since both states had weekly case counts that were higher than expected (greater than one standard deviation above the mean) in advance of the peak week. These signal weeks occurred before the case count for the year, exceeded the total case count for the previous year, and would have been a warning to emerging increases in the disease. This result does not indicate a clear advantage to using the news for disease detection or situational awareness, and makes a case for not adding an additional information source for decision makers due to no proven utility.

Google Search Frequency

In the search query data for California, Michigan, and Ohio, there are some notable spikes in search frequency. It was critical to determine if these spikes are “true” (reflecting increased cases in the population searching) or “spurious” (not reflective of health impacts in the population being analyzed) as defined by Chan (21). None of the spikes met the definition of "spurious" (defined as five standard deviations above the mean of the week of the spike plus the four previous weeks), meaning each could likely be caused by normal disease transmission (21). Although Breyer suggested that values greater than three

standard deviations from the mean may be excluded since they were likely the result of nonmedical events (34), our findings indicate that three standard deviations were indicative of “true” positives. This finding may be due to the limitation of the data insomuch as the maximum is a relative volume that cannot exceed 100 regardless of what the true value is.

These findings do not mean definitively that the spikes have no relation to a news story or some other external factor such as announcements of potential new treatments, vaccination, corporate interests, and/or celebrity cases. It is possible that spam and popular news articles containing key phrases can influence search results and create peaks in activity that may not be reflective of the incident being researched (23). For instance, Walgreens began marketing the availability of pertussis vaccinations at over 150 California stores on September 17th (107). During the two weeks directly following that announcement, search frequency for pertussis was more than one standard deviation above the mean in California. Therefore, it is unclear as to whether the increased searching is the result of a corporate marketing campaign raising questions in a community, instead of pertussis cases. Fortunately, there were no stories of a celebrity suffering from pertussis during the course of this study period, which would have been a potential confounder.

If an individual searches for information online, they have ‘adopted’ it. Their adoption is motivated by either ‘rational’ action where a person decides independently to search in response to external information (news story, announcement, rumor, etc.) or because of ‘social transmission’ where the person searches because other people are searching (84). It is not possible to differentiate between the two reasons for the search. Events like “bird flu” and fears associated with it were abrupt and authoritative, and these types of events can change the process from social to independent decision (84). Fear or curiosity may result in panic-induced searching, but some of this searching may be due to legitimate health risks (21). With search terms, it is not possible to know the user’s intent or motivation (76), and

each individual search could be for a relative, because of a news story, or for preventive purposes. One example is clinicians, who may be using search engines to gather available information for patients from the web (78), which has the potential to skew or inflate data, if both patient and practitioner are looking. Additionally, incorrect self-diagnosis may result in search terms being utilized that are not clinically accurate (21). Another consideration is that multiple searches done by the same person are considered independent, so you could have a small population within a region that searches the same terms on a daily (or other frequency) basis, and each search would count as an independent event despite being the same individual or household constantly searching for new or updated information.

Google Trends approaches may contain content inaccuracies for a number of reasons, including data sampling issues as well as the approximation methods used to compute results (94). Google Trends data is based on a sample of Google web searches that may result in a non-representative sampling bias (78), and the results are dependent on several assumptions and approximations that may obscure or misrepresent true trends in search traffic (20). Search information is not completely transparent because Google internally set a threshold below which data trends are not released or identified.

Per Carneiro's work, search frequency data may not be useful for diseases with low prevalence (not significant search volume) (20), but there was no specificity on what she would consider as "low prevalence." The pertussis data collected for this work indicates that pertussis (despite its relatively low prevalence in relation to other diseases), had sufficient search volume for spikes and variation to be identified.

Google Search Frequency and NNDSS Comparison

The sensitivity of Google search frequency was relatively low in all three states, indicating there were many "false positives" identified if Google was used as an indicator of an actual spike. For both sensitivity and specificity calculations for Google search

frequency, it may be have been too simplistic to look at the week of the search spike against the week of the case count spikes since Google was shown to be an earlier indicator. The methodology chosen assumed that Google search frequency would spike at the same weeks of case reporting spikes; however, the known incubation period of the disease, as well as reporting lags, may have been a potential reason to complete an assessment where both incubation period and reporting delays were incorporated into the analyses. It might be important to recalculate this analysis using some sort of re-calculated "Google week" (i.e., Google week plus incubation time) for a more effective understanding of sensitivity and specificity. This type of approach was not identified in any of the reviewed literature, and this may be because of its potentially limited statistical rigor.

The low accuracy values that resulted from calculating the F-1 scores may also be indicative of this potential time lag between people searching for terms, the incubation period, and/or reporting delays. There are potentially temporal delays that should be considered when assessing the utility of these novel sources of data based on human behavior and practice. The slow pace of each step of the medical process likely varies by provider, treatment facility, and laboratory procedures; therefore, no known standard can be applied.

Lastly, both NNDSS and Google Insights used the same date to start and end weeks throughout calendar year 2010. This dating improves the validity of the comparison between the same data set because the timing matches perfectly.

Social Media

One of the most significant findings in this research was the extremely low proportion of Tweets that were considered relevant coming from the original data set (<.06%), and although the researcher cannot be sure that there was a sufficiently sized data set (23), the ratio of relevant to irrelevant supports the conclusions that there were no findings from

Twitter data regardless of the raw Tweet count. While Gupta found that only 17% of the Tweets containing information about an event were deemed to include credible content (46), this work found that credibility is nearly impossible to confirm when assessing personal health information.

While perceived severity and intense news coverage are likely factors that have been shown to dictate Tweet posting activity (23), it is possible that these findings indicate the opposite is also true. Low news interest and lack of awareness by the public may result in decreased postings to Twitter.

Those who Tweet about health issues may not be representative of the Twitter population, and the Twitter population may not be representative of the general population (23). It is also unclear how representative attitudes and opinions expressed through social network sites are of attitudes and opinions of the broader United States public, and the differences may limit generalizability (87). There is disproportionate Tweeting amongst Twitter users, with 22% of users accounting for over 90% of Tweets (108). This distribution could indicate that, regardless of the similarity between users of a site compared to the total population, the more indicative (and unobserved) variable is that which differentiates those who post to social media sites as compared to those who have accounts but choose not to add content.

While crowds are prone to add their personal opinions to data, and opinions overshadow the credible content (50), these findings did not indicate a likely crowd presence in shaping information in this context.

Social Media and NNDSS Comparison

The extremely low number of Tweets that were relevant to pertussis limited the ability of the researcher to compare epidemic curves, timing, or utility of social media (namely Twitter) against official reporting. This may be the result of analyzing a disease that

has a very low overall incidence and for which mostly young children are affected. These findings are important though, because they signal that while social media may have value for more common diseases, its utility may vary by disease and/or population at risk. Unlike the findings of Kwak, only a very small number of Tweets were related to news content regarding pertussis (109), which may be a result of the high number of irrelevant Tweets due to the queries used.

Combination of Infodemiology Sources and NNDSS Data

Timeliness

Recognizing that early indicators mean that the data source "detected" an event (110), this research then shows that infodemiology sources, specifically news media, internet search patterns, and (marginally) Twitter, are detectors. Characteristics that influence detection of outbreaks include magnitude of the signal, shape of the signal, and timing of the outbreak (2), and the data here have those parallel characteristics. The incubation period for pertussis is twenty days (3), so any tool that provided information less than twenty days in advance of the official reporting would be critical to prevent morbidity and mortality. Google search patterns did meet this twenty day timeline for identifying the peak four to eleven weeks in advance (twenty-eight to seventy-seven days). The same was true for social media, where Tweets provided an indicator fourteen to twenty-six weeks in advance. News media did not meet this twenty-day threshold to be within the incubation period. In comparing NNDSS data to infodemiology sources, it may have also been important to assess if the alternate sources identified either the initial (primary) spike and/or the secondary spike. This work focused only on the initial spike in the curve.

Sensitivity and Specificity

When looking at the accuracy of all three infodemiology sources in combination, one key finding in California was that there were two weeks where both Google and Twitter

indicated a spike, but there was not a spike in NNDSS data (creating potential false positives). This inconsistency has the potential to mislead decision makers if they were working under the assumption that the signal of multiple novel sources is stronger than the signal of a single source. In both Michigan and Ohio, none of the infodemiology sources showed a signal at the same week as any other week indicating that each source has unique value for determining indications of increases. The purpose of event detection is to differentiate rapidly between baseline occurrence of a disease and more severe outbreaks with high accuracy when the outcome of the event is not yet known (2); therefore, identifying data sources that can accurately identify true positives before the event as a clearly identified epidemic would be a critical capability.

When assessing sensitivity for each source independently, or the ability of the infodemiology source to identify a true positive, none of the sources had significant success. Alternatively, the novel information sources each had high specificity (ability to identify true negatives or actual weeks with no increases in case count). In calculating the F-1 score for each infodemiology source, the researcher was attempting to assess the accuracy of these alternate sources. Neither news media, nor internet search patterns, nor Twitter provided accurate information as compared to NNDSS data.

Correlation

Neither news articles nor Twitter data could be correlated to NNDSS, and the correlation between internet search patterns and NNDSS only existed between Michigan and Ohio. This discrepancy may be indicative of variation in population size and internet prevalence amongst these states. These statistical findings may not align with visual inspection of the temporal graphs for any of the three states, so it may be worthwhile reviewing the utility of correlation analysis when assessing infodemiology data. There also may be inherent reporting biases in any/all of the sources due to heightened surveillance for

certain diseases or in regions where there is a predisposition for disease activity (67), so it is possible that once news reports started covering pertussis that activity resulted in an uptick of Google searches and Tweeting. There is also the potential for political change to have an impact on the findings. Media reports can confound and influence infodemiology data, because of an “epidemic of fear” (1).

The epidemic curve of these events may influence detection and findings. Characteristics of the disease process (such as duration of the incubation period) influence sensitivity and timeliness of outbreak detection, with fast outbreaks being detected more often and more quickly than slow outbreaks (2). Therefore, findings from this pertussis work may not be applicable to epidemiologically diverse conditions. The calculation of predictive values assumes that all the weekly data are both homogenous and equally likely to be an epidemic week (110), which ignores the potential for seasonality, school year, and other such realities that impact disease transmission; therefore, potentially providing insight as to the results obtained in this research.

Survey Results

The survey results suggested that when looking across all phases of situational awareness, identifying early indicators of disease was the phase in which infodemiology sources were seen as most compared to the strengthening comprehension or informing prediction phases. This aligns with past research indicating that early detection was the area of most value for novel information sources and that current methodology (like epidemiology investigations) would be more useful in characterizing and comprehending the specifics of an identified outbreak.

Social media was the infodemiology source that had the highest value for "Unsure" responses in comparison to news media or internet search across all three phases of situational awareness. Moreover, social media was deemed "Not Valuable" more than other

sources for identifying early signals and strengthening comprehension. This aligns with findings of skepticism about social media amongst professionals, and it indicates how both news media and internet search have some undefined characteristics that make the sources more believable as compared to social media. In 2001, both the United States government and the WHO had already identified the value of open source, e.g. via Henderson endorsing ProMED-Mail (15) and WHO using GPHIN's newspaper-based content (18). In the past twelve years, the public health community has added internet search into the toolbox of sources to use (in the absence of an official proclamation by professional societies or authorities about its value). The current question is if (or when) social media will gain such perceived value amongst the professional community.

Additionally, survey respondents had mixed opinions over whether it would be possible to have "too much information" during a disease outbreak. This finding indicates that public health professionals may be more informed than the general public over whether "more is always better" when it comes to information (57). It is possible that public health professionals have had more exposure to information overload than some other professions.

One of the biggest potential limitations to the survey is coverage error since the study population is restricted to members of a single professional organization (ISDS). Amongst those who are part of the study population, there may be variation in attitudes or beliefs between responders and non-responders of the survey, resulting in selective nonresponse error. It could be theorized that those who do not believe or perceive utility in open source information may choose not to take the survey, causing a potential over-estimate of utility or value. Nonresponse error could also occur at the item or question level if some questions are perceived negatively or are not clear to the participants. The survey response rate was also relatively low (13%) as compared to the average response rate of 34% found in a meta-analysis of web-based surveys by Shih (111). This limitation may not

be significant if respondents were representative of the intended participants, but the researcher did not do analysis comparing responders to non-responders. Additionally, it is unclear amongst the survey participants if they have a decision-maker role within their organizations, and this may have been a critical question to include in order to understand if the respondent does (or does not) have to make critical decisions during a disease outbreak. There is also the potential for measurement error, in the event that some of the respondents answer the question inaccurately due to confusion about the question intent or terminology.

Boundaries of Research

There are a number of summary points that can be made about this research and its' future value. These are limitations caused by the novelty of the subject as the topic for investigation. There are few established pathways to assess infodemiology and the investigator had to overcome multiple barriers to achieve what has been described in the study.

- This research only focused on a single disease [pertussis] and a single official data set [NNDSS] over a single year [2010].
- Actual decision-making was not utilized in evaluations of information utility.
- Only content (news articles, Tweets, search terms) written in English was used.
- No raw search information was used; only publicly available Google Insight or Google Trends content was used
- Traditional demographic data from surveillance systems were not used.
- Information sharing Tweets and news stories were excluded, since they do not provide insight about actual personalized health conditions or concerns.
- Multiple reports of the same event were excluded to prevent popular posts from saturating the sample (23), hence the need to determine a subset of news stories as

novel. This multiplicity included retweets ("RT") and Associated Press (or other news service) stories repeated in multiple online newspapers.

- Utilization differences and reporting differences amongst different mediums was not assessed, and socio-demographic and cultural variations were not included in this analysis.
- Information about public messaging campaigns and marketing campaigns was not included in this assessment, despite the potential for marketing or promotional activity about the disease being studied, related prevention measures, or available treatments to change the level of activity on any of the various mediums being studied.

Limitations in Summary

In addition to the limitations addressed within the Discussion section for each data source, there are some overall limitations of this research that should be noted. Only a subset of all-source epidemic intelligence data was captured for this assessment, although the data types covered the variability of content types available to a public health department today. This research only looks at a single type of event in a limited geographic area; therefore, it may not be generalizable to other geographies or diseases. Generalizability may also be limited due to internet access that may vary by socioeconomic, demographic, and geographic subpopulations (31). There may be a selection bias because of those who use (or do not use) the tools studied (Google search and Twitter).

Known technological challenges exist when trying to utilize infodemiology for epidemic intelligence to include topic detection and data acquisition from high-volume streams, data characterization, categorization, and information extraction (104). Information overload may result from the large amount of data being collected, and it may be challenging to distinguish the signal from noise that, in turn, may decrease utility (9).

Alternatively, ensuring capture of all relevant data is not possible, so useful content may be missed or not appropriately classified as relevant. The list of search terms used for all of the data sources (see Appendix B) may not be complete, and some information related to pertussis may not have been captured (70). This list may not have fully taken into account how people use terms differently depending on cultural and language backgrounds as well as level of education (20). Additionally, it may be challenging to apply consistent exclusion criteria when extracting data for analysis and translating that content into data for each variable (67).

Information quality is dependent upon accuracy of self-reported details and sufficient numbers of people reporting, so it may be prone to reporting bias (22). False information (mis- or disinformation) may have resulted due to using citizens as reporters, which may have caused reporting bias (9). Quality concerns also exist around geographic data, which are extracted from the internet protocol address and may not always represent accurate location information (30).

Factors like news events, substance abuse, and cultural differences might influence behaviors (33) and reporting, both incrementally and over longer periods. Not everyone who gets sick actually goes to a health care provider, and not everyone who goes to a health care provider is accurately diagnosed. Not everyone who gets sick Tweets about it, reports it, looks up their symptoms on a search engine, etc. Seasonal variation is also a potential for any of the data sets, and it was not addressed here due to only a single year of data being analyzed.

CHAPTER 8. PLAN FOR CHANGE

There is an agreed upon need for improving public health situational awareness, which has been documented in various policy documents and regulations. Unfortunately, significant progress has not been made on the development of policies, procedures, tools, or methods to implement these improvements at the Federal or local public health jurisdictional level. More specifically, a comprehensive approach to incorporating infodemiology content, including methodology, protocols, and business practices needs developing into a coordinated situational awareness program. This program would include training for leaders that is focused on interpretation of new media (48). Many of the survey respondents identified areas where there were knowledge gaps amongst public health professionals. As a way to catalyze this, I am proposing the development of a *framework* for implementing infodemiology as a content source for public health situational awareness. The *framework* can help facilitate eventual adoption of expanded surveillance practices. The term framework is used in the place of "plan" since this is more of an outline of general goals and processes that can inform a plan for implementation. The policy application of this work may require inclusion of new information sources into existing reporting pathways within public health response organizations to improve situational awareness, which would improve efficacy of response.

Current State of Policy and Practice

Currently, there are a number of policy documents that call for development of improved situational awareness strategies, systems, and tools such as: the National Health Security Strategy (NHSS), the National Biosurveillance Strategy for Human Health

(NBSHH), and the National Strategy for Biosurveillance (NSB). The earliest was part of the Pandemic and All-Hazards Preparedness Act (PAHPA) of 2006, a mandate that the US Department of Health and Human Services shall establish “near real-time electronic nationwide public health situational awareness capability to enhance early detection of, rapid response to, and management of, potentially catastrophic infectious disease outbreaks and other public health emergencies that originate domestically or abroad” (112). More recently, the National Security Staff led the development of the NSB, which reiterates need for "the identification, sharing, and integration of essential information to expedite incident detection and assessment" (113) . Some of these documents do make reference to evaluating and integrating novel information sources such as news media and "web-based social networks" (114) or incorporating social media as a "force multiplier" (113), but there is no direction or guidance on how to include this content.

When it comes to the actual practice of using infodemiology content to improve public health situational awareness, a few informal examples exist at the Federal level, but there is no documented position or endorsement. Also, a small number of public health organizations at the state or local level have cited use of novel sources, but examples of infodemiology in practice are rare (as are stories of public health situational awareness). When looking at global public health colleagues, the WHO is able to use information on suspected outbreaks from unofficial channels (18), and the United States should do so as well.

As global travel and trade continues to increase, the importance of identifying new and better methods of detecting emerging disease will continue to grow. Additionally, the constraints that exist on the reporting of emerging infectious diseases are unlikely to disappear over time. These include: fear of repercussions on trade and tourism, delays in government clearing and information sharing, tendencies to err on the conservative side,

and an inadequately functioning or non-existent surveillance infrastructure (104). In the case of constrained reporting through traditional surveillance sources, or the disruption of these sources due to a catastrophic event, the utility of these nontraditional sources may be more evident.

When reviewing technologies and tools used in public health, no toolset currently exists in the emergency management or public health community that blends official content with social media (115). In addition, as public health continues to face dramatic funding cuts, methods to improve surveillance using free or low-cost data is important. It is incumbent on the field to consider alternate sources, which may have the potential to reduce the amount of work required of over-burdened public health professionals.

Identified Need

While the desire for improved situational awareness exists, very few specifics on how to accomplish this improvement have been proposed. The definitions for public health situational awareness vary and do not clearly define it as a capability, a tool, or a construct that *must* exist. This dissertation has argued that improved situational awareness leads to improved decisions. So, to improve public health outcomes during disasters, it is necessary to improve situational awareness so those decisions are as informed as possible. This need is especially critical at a time of limited resources, when decisions must maximize effectiveness with minimal drain on existing resources.

Situational awareness must be as timely and accurate as possible. Using non-traditional data evidenced in the findings outlined in the preceding chapter is one way to improve real-time public health situational awareness. This novel content fused with other information could alert public health officials at earlier phases of an outbreak, enabling them to rapidly respond and to facilitate case containment, epidemic investigation and access to treatment to reduce morbidity and mortality. The analysis described in previous chapters

showed this value from a data-driven perspective. Similarly, the survey indicated the positive attitudes of professionals who would be users of this content. Success in containing an emerging epidemic is dependent upon rapid identification of the original case cluster and continuous detection of a high proportion of ongoing cases, preferably within 48 hours of a case arising (116). In addition, rapid containment leading to quicker prevention and control, it also results in more advanced warnings to the public and additional time to prepare countermeasures to decrease morbidity. In a time of limited resources, it is more critical than ever to know what actions to take and where to act as close to real-time as possible where incidents are constantly evolving (48).

Many recent disasters have shown that social networks are a growing way for critical and accurate information to be shared amongst the public, especially information that would be otherwise hard to obtain (117). In some organizations, leadership may not be passionate about the use of social media, believing that it is a passing fad that should be ignored (115); therefore, a document that trusted professionals developed, citing known successful uses may help to reduce some skepticism. Both at-risk populations and traditional media are adopting and considering social media as a viable information source, and indicate its growing value to communities overall (117).

Recent events, like Superstorm Sandy, further illustrate the need to include these kinds of data in the work of disaster response organizations. Consider that if only .001% of the over twenty million Tweets from the first five days of Sandy were actionable, and even if only half were accurate, this would still mean over one thousand informative, real-time tweets (over 25 pages) providing relevant, actionable, timely information (118).

Filling the Gap – Developing an Implementation Framework

While a need has been identified, there has not yet been any policy implementation or guidance on how information that can improve situational awareness (like infodemiology)

should be incorporated and utilized. The first critical step will be the development of a *framework* for utilizing novel data sources and fusing these sources within any given situational awareness approach or tool. While the ongoing discussions amongst the federal public health community, as well as state, local, and territorial health organizations continues regarding what situational awareness is and how it's accomplished in the broadest sense, the findings in this research indicate that there is are valuable new information source to include. The intent of this implementation *framework* is that it can easily fit into any broader situational awareness approaches or plans, while providing a resource and methodology to facilitate incorporating these data sources.

For the implementation *framework* to be considered useful, it must be developed as a coordinated effort amongst key stakeholders. This group can develop a consensus and identify agreed-upon principles or concepts for infodemiology data, and may include professional organizations (such as the International Society for Disease Surveillance), policy makers, practitioners, and technology partners. The process would also require a champion to ensure that the work continues to evolve and remains a focus area, this might be a White House representative or high-level official in a key department.

Several emerging and intersecting policy streams or interests at the federal level suggest a window of opportunity to convene a group of stakeholders to develop the *framework*, especially in the space of utilization of news media, Internet search patterns, and social media content. Increasing discussions of transparency and "big data" seem likely to result in an increased desire by elected staff to use publicly available data to improve government activities as was done in recent presidential campaigns. The ever-evolving availability of information and the tools used by the public would require this to become an ongoing discussion, with the *framework* as the first step in this dialogue.

In addition to the active inclusion and participation from stakeholders, it will be critical to inform stakeholders throughout the development process. Communication will need to be broader and include more than just the stakeholders convened for the *framework* development in order to reduce the likelihood of missing any key parties. Public communication will also be critical since the data sources being discussed are likely to concern the public and interest groups concerned about privacy and the role of the government in viewing citizen-generated content.

Elements of the Framework

When developing the framework, it became obvious that any guiding document must effectively address various stakeholder groups. For this work, the stakeholders are: policy makers who will need to write strategies and legislation related to infodemiology for situational awareness, the decision makers who may adopt infodemiology into their daily work for disease detection, and the operators who will do both analysis and development of products that decision makers would use. While policy makers serve as strategic leaders, the decision makers lead operational aspects, and analysts provide support at a more tactical level. To address these various communities, the recommendations below are focused on each community identified.

For Policy Makers

The creation of policy that incorporates infodemiology for public health situational awareness will inform how and when decision makers utilize this new content source and identify required data or tool development which staff within public health organizations will need to do.

The purpose of incorporating infodemiology is to reduce time for event detection and increase accuracy of outbreak verification. Infodemiology data has the potential to provide additional insight and inform or strengthen early indications of an

emerging outbreak. Survey respondents submitted that infodemiology content provides the most value in early indications of a disease outbreak. Specifically, the focus should be on how Step 11 (indicated in Figure 8.2), which could happen sooner in time if public health officials changed their approach to situational awareness to incorporate additional data types. If the duration for each of intervals (indicated as arrows) could be reduced due to additional data to facilitate determination of a disease, this might decrease the impact and reduce the spread. Steps 1-7 are the responsibility of state public health, while Steps 8-11 are the responsibility of the Federal government. Infodemiology could improve timeliness for both types of organizations.

Figure 8.1. *Modification of Jajosky's sequence of gathering and using health-related information for public health (13). Each arrow represents an interval; each square represents a step in the process.*

Also, adding new data sources, specifically infodemiology, can improve accuracy for verifying the specifics of an outbreak. According to Heymann, there are four phases to outbreak alert and response: collection of information about potential outbreaks, verification of outbreaks, communication to partners/others, and containment/coordination (18).

Infodemiology data can be especially valuable in expediting and/or improving verification of outbreaks, as has been shown in the previous chapters.

Think about broader applications of these data to public health issues beyond disease surveillance. Additionally, this type of work may indicate behavioral and risk factors that exist within communities in place of more traditional phone surveys, such as has been shown for tobacco use where Twitter data regarding smoking aligns more closely with tobacco usage than survey results from the CDC (43). Nontraditional correlations may also be identified, spurring research into potential causes or impacts, such as correlations between mood and obesity (43). Additionally, this type of information could be useful to assess mental health after a disaster since people may not be comfortable coming forward (like suicide). With current surveillance and data collection methods, by the time a trend of increasing suicide risk is identified, it may be have already peaked and waned, making it too late for an intervention (85).

Developing a better understanding of the utility and applicability of these data and information can inform or be applied to established models that characterize social versus independent interest in health issues (84). This type of content could be used to inform policy makers on the effects of their recent legislation or policy in a timely way (31), as well as improve planning, monitoring, and intervention. Changes in human behavior in response to a disease outbreak affect the disease's progression (119), so it is important to understand how and what information is reported in news and social media sources so that behavioral changes can be anticipated and planned for. People’s behavior changes because of awareness of the presence of a disease, which indicates that their postings on social and news media sites may reflect awareness of a disease and not direct observation (119). This behavioral shift is a potential area for confusion in public messaging and disease detection. Because data sources are likely to facilitate outbreak detection when the data is collected

frequently for a large proportion of the population and when the data relates closely to disease incidence (2), these sources are becoming key resources. These data are also becoming key resources because news and social media are becoming more representative of the population, and can provide richer context for ongoing events.

Consider the potential utility of nontraditional sources as a feedback loop that could be used to monitor public perceptions, concerns, and behaviors. This feedback loop could also be used to examine and evaluate effectiveness of knowledge translation strategies and tailor future communications (23). Content and sentiment analysis can enable public health departments to rapidly identify real and perceived concerns and issues being raised by the public. During the early phases of an outbreak, data about the public's response and behaviors would help inform a more effective public information campaign (42). Recognition that the potential for rumor spreading via new mediums will exist, and there is need to develop policy around how and when these sources should be utilized and what the steps are for verification of content.

Studies have shown that animals will respond spontaneously to those who have useful information, as shown by studies regarding location of food and migration routes (120), and likely about threats or dangers in the way lemmings follow others. If these findings apply to humans, then it is critical for public health practitioners to listen to those in the social web space since the population as a whole may listen and respond to few individuals who are speaking out. Studies of social learning amongst animal populations indicate that the larger a group is, the smaller the proportion of individuals needed to guide the group (120). Because of this, only a minority of the population need to do something for it to lead an entire group in a cohesive direction (120) and this may apply to utilization of new web tools for networking or posting, where a small population could move a whole community towards using or posting content.

Recognize the public perceptions and expectations around these data types. The general population feels that local (70%) and national (80%) emergency response organizations should regularly monitor social media sites and promptly respond (25). This monitoring will require public health agencies to make it clear to the public what aspects, and for what purposes they are monitoring social sites to reduce failed expectations.

For Decision Makers

Public health officials serve as decision authorities, especially during a disaster or crisis. These professionals are bound by policy developed for them to follow, while utilizing information provided by analytical and support staff. Due to this unique role, it is critical to incorporate some framework elements focused on these individuals.

Include approaches to reduce this burden of information overload, since infodemiology would be an added data burden. Infodemiology content may make public health officials' jobs more challenging by introducing insight about additional outbreaks of low impact conditions, which may distract them from higher priority events due to media (9) or society focus. For rapid analysis of large amounts of data (of the type public health officials are dealing with daily), three activities must occur in tandem: expert analysis, crowdsourcing, and machine learning (50). Experts are needed to rapidly generate hypothesis for emerging events as well as to validate, confirm, deny, or characterize outbreaks as they are occurring (50), and public health officials must always make the effort to include epidemiologists in these steps rather than relying solely on machines. Multiple experts across various domains may be necessary to obtain an accurate depiction of a situation that is unfolding (50). An algorithm can present these novel data sources to an expert in the form of an indication, and the expert can declare or confirm the outbreak (50); however, those algorithms first need to be developed and made available.

Infodemiology should be seen as a data source to complement or augment rather than replace traditional information sources. The initial need is for the development of a system that maintains accountability of sources as well as control for users regarding content included or excluded, while still incorporating content that is not "official" (115). Survey respondents were eager to better understand the value add that these sources provide in comparison to existing information streams. There is a need for "fusion" of social media content with existing disaster content and business processes to enable effective crisis informatics (51). The question of how to deal with unverified reports remains, but the idea of utilizing sources together indicates a new way to inform and improve reliability of any one source. Integration of multiple approaches (to include both traditional and infodemiology content) offers the greatest promise for the future of public health surveillance (9). New, alternative sources can fill gaps in current approaches, but not without inclusion of traditional epidemiologic approaches (9), and currently no policy about these combined uses has been developed.

Consider which diseases or events infodemiology have a high potential to reduce morbidity and mortality and focus the plan on that sub-set. Incubation periods can be used as a surrogate measure for communicability timing, so if any of these methods provided reporting that indicated cases one or more incubation periods before official reporting, then they should be utilized to facilitate prevention and control methods (13). If the plan shows value and works for this subset, the approaches can be expanded to other diseases or events of interest and this will reduce resource demand.

Recognize the challenges and limitations of these data sources. Self-reporting lacks diagnosis confirmation, which can cause challenges for validation, filtering, and public health interpretation (10). Redundant reports (multiple sources may report the same event) may result in misclassification or overestimation of impact. “Worried well” may exaggerate

impact of a health condition in one or more of the studied regions. It is necessary to include plans to monitor for, and mitigate against these potential challenge areas.

For Analysts and Technologists

Analysts and information technology professionals serve in a support role to public health officials, providing reports or products to support decision making. These staff members of public health departments are the implementers of policy that has been developed and as well as the people tasked with evolving available products and tools to keep leadership informed.

Constantly evaluate both new and existing infodemiology sources and methods. As a first step, this includes assessing strengths or weaknesses of infodemiology (and all) data sources, including traditional reporting, to look at timeliness and accuracy. This would be especially valuable to public health officials in information–poor regions. In order to collect, monitor, synthesize, and fuse information, officials must be able to identify critical utility of emerging sources. Also, analysts must assess the value of the information from infodemiology sources before incorporating it. The cost of information in comparison to the value of information collection should be taken into consideration (121). If the collection will be resource-intensive, determine if there are other sources that can be reduced or eliminated since funding is unlikely to increase for public health.

Ensure whatever methods or approaches taken are flexible and open to new technology. As technology continues to evolve at a rapid pace, our approaches to utilizing the resultant information from new technology must be extremely flexible. There are ongoing changes in the information types, structures, and tools, and the *framework* must include an adaptable approach to incorporating emerging capabilities. One potential emerging concept would be incorporation of citizen-generated reports of events directly to

public health authorities. The question is whether the public would participate in a "see something, say something" campaign like that used by New York's Metropolitan Transportation Authority (45) if the focus were on contagious diseases. Would the public participate in reducing the spread of disease by reporting if people around you are coughing more? This public reporting would work as a neighborhood watch applied to expand situational awareness, giving citizens the capacity to directly communicate emerging events to officials (45). If so, the development of citizen-generated reporting capabilities needs to be semi-structured (so that people can report what they want), real-time, open, geo-aware, and accessible across multiple platforms (50).

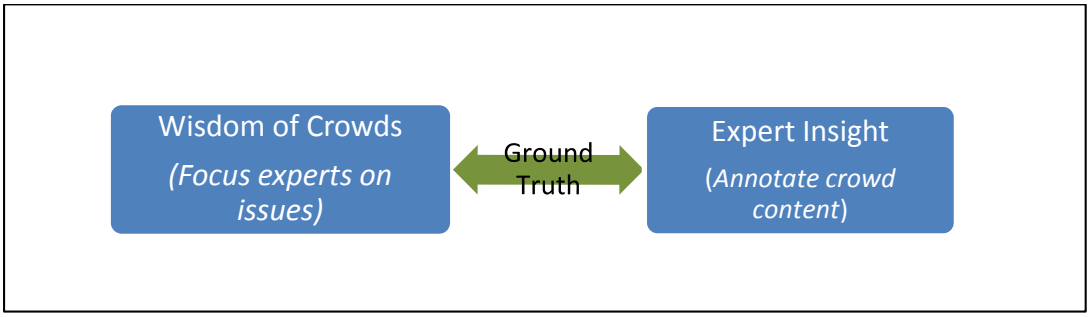
Include approaches to rapidly identify false or misleading content, potentially via data quality methods. Public officials will need to be aware of the potential for malicious information or misinformation being distributed and develop ways to mitigate those attempts and to reduce their impact (48). These attempts may lead to criminal actions against people who purposefully distribute misinformation on social media channels (48).

Include an approach or methodology for setting of thresholds for these alternate data sources, like those used with traditional data sources. A threshold that is too low will result in many false alarms, but will enable the earliest possible detection. A threshold that is too high will miss critical early indications of the disease of interest. The necessary timeliness for reporting should be assessed based on potential decisions (and even each disease's incubation period) to set appropriate thresholds.

Effectively communicate the validity and reliability of infodemiology sources to public health leaders, especially in times of rapid decision-making. The survey findings support the conclusion that mechanisms must be developed so personnel have awareness of the sources involved in any informational product or visualization as well as

their respective trustworthiness and validity. This policy will ensure that the level of confidence (perhaps in the form of a reputation score) is made known to the decision maker (50), as well as to allow the relationship in Figure 8.2 to be better understood.

Figure 8.2. *Adaption of Kamel Boulos' Integration of experts, crowds, and algorithms (50)*



Action Steps to Facilitate the Framework's Ability to Improve Public Health

1) Publish in professional journal(s) and present at professional meetings on what the *framework* is and why it is a critical step in improving public health situational awareness.

2) Find advocates in leadership roles, to both continue the development of a broader public health situational awareness approach, as well as to ensure inclusion of infodemiology sources in any strategies or tools developed.

3) Advocate for infodemiology content to be included when funding opportunities (grants) include public health situational awareness, in conjunction with the framework developed.

4) Make recommendations for infodemiology concepts and utility to be incorporated in training, along with key elements of the *framework* (via the Association of Schools of Public health) and/or continuing education credits for Certified in Public Health (CPH) colleagues or other professional certifications.

5) After the *framework* has been released, conduct an impact evaluation to assess if the *framework* did increase the number of public health organizations that utilized infodemiology. This impact could be assessed as either incorporation of the *framework* document into an evolving situational awareness strategy or plan, or it could be an independent operational assessment of how many agencies used infodemiology after the *framework* as compared to prior.

Further Research

Potential areas for further research have already been identified as gaps and identified in the published research. To enable robust future research, findings from efforts such as this dissertation need to be published in journals. Research will need to continue as the medium and amount of social media usage changes over time within the overall population and within sub-populations. There are research findings that indicate that both news and social media have some capacity to provide unique information and insight that is not available through more traditional public health data sources. The utility and value of this data varies both by condition or disease, and by source, and needs to be examined further to understand fully the maximums.

To enable future researchers to contribute, as well as to facilitate practical application of this work, there is a need to build a freely available public health search term vocabulary for each medium, using natural language processing. For each of the media, terms need to be evaluated both individually and as symptom complexes or syndromes (20) to assess utility. These information sources are often unstructured and difficult to interpret, requiring advanced computational resources to implement effective categorical or quantitative assessments. Increased research on natural language processing and the development of related tools for information retrieval, text classification, and text mining are crucial next steps for converting text to structured event data. Additionally, methods like

Latent Dirichlet Allocation (LDA) need to be used for identifying new terms for disease topics that are not directly intuitive but likely relevant (122). Specifically, this type of methodology would allow for new slang terms for diseases or symptoms to be rapidly identified and incorporated into analyst dashboards. LDA has been shown to extract valuable topics from large amounts of data (122), including user profiles using labeled LDA (or L-LDA) (123), and this capacity will become more critical as the amount of information available continues to grow astronomically.

Validations of findings would need to be conducted in comparison to traditional, resource intense, observational or cohort studies (31). Investigation of the key characteristics of an effective surveillance system (representativeness of system, outbreak detection algorithms in use by the system, and specificity of the algorithms) (2) has not yet occurred and should be done for each of these novel surveillance sources. Both sensitivity and specificity are unclear and false positives have the potential to increase workload on already overburdened public health employees. One survey respondent indicated the critical need for a real-time validation effort, rather than retrospective analysis. This would reduce the potential risk of retrofitting data based on historical insights and also provide an improved understanding of how this content would work in daily practice.

Many survey respondents indicated the need for improved baseline data for these emerging sources to distinguish threatening anomalous events more precisely and to understand more clearly the background reporting, both on a regional and per disease basis. This differentiation is especially critical in a time where the amount of information available to public health officials and everyone is growing at an astounding rate; every two days there is more information created than the amount between the dawn of civilization and 2003 (124). Multivariable linear regression can be used to predict the normalized frequencies for each of these mediums based on factors like season, schools in session

(77) or some similar type of “control analysis” for various regions or periods. Improvements in aberration detection algorithms to train Bayesian classifiers to increase positive predictive values and therefore decrease false alarms are also important. There is a need to develop methods to decrease false alarms and to vary the weight of different sources of information, as well as identify ways to distinguish events as high confidence due to multiple sources with corroborating information.

Better understanding of health behaviors and concerns amongst the public could be assessed from this type of data (30), although this understanding would require development of new methods as well as comparison studies (as mentioned above) to be conducted. For instance, new symptoms or home remedy treatments may be first identified through these alternative information sources (125). Twitter is potentially suitable for longitudinal text mining (to identify changes in opinions or responses) and can provide instantaneous snapshots of the public’s opinions and behavioral responses (23). It is reasonable to presume that other new or emerging technologies may have a similar value.

The next area of research is needed in determining policy implications of these findings: How much of a change from baseline will warrant further investigation or deployment of resources for investigation (83) or engagement—i.e., what is “actionable” in this space? It has been shown that the spread of information about a disease has the potential to impart benefits and reduce the spread of disease because information creates awareness, and awareness triggers the tendency towards protective behavior (119); however, the dynamics of behavior because of social media has not been fully examined. There is a need to develop methods to assess the impact that disease or health-related messaging (to include rumors) spreads through these new mediums, as well as to understand factors that amplify the information spread.

Conclusions

What We Know Now

This research showed the potential utility of timeliness for early detection of pertussis in two of the states identified. The work also showed the potential for news or Twitter to serve as a leading indicator when compared to traditional data. These findings may or may not have generalizability to other diseases or other places, or even to current detection given the rapid evolution and availability of internet technology. The research did indicate that it is unlikely for all three novel sources to give a clear and obvious signal in advance of an outbreak, but maybe just one or two sources might point to a potential outbreak. Previous research findings proposed that one can be most confident in the synergy of infodemiology approaches when combined with more traditional syndromic or laboratory surveillance, and rigorous evaluation of this combinatorial approach has only just begun. With that reality in mind, public health must develop a *framework* on how (and when) to use infodemiology content while this research continues.

What We Need to Do

In order to act on the findings of this research and the previous findings, we need to acknowledge that incorporating new content types into any organization is not the role of one office or function, but rather it is a cooperative effort amongst all elements. Although all the actions listed in Table 8.1 are listed under a single office, we recognize not all these functions are solely the responsibility of any three groups, and most items in the chart are crosscutting due to the collaborative nature. This is especially clear when looking at the fourth item for each decision makers and analysts, where it is clear that knowing limitations and informing about limitations are things that cannot be done independently.

***Table 8.1**. Summary of activities to be undertaken by stakeholder groups to implement framework*

Policy Makers	Decision Makers	Analysts & Technologists
Infodemiology reduces time & increases accuracy	Infodemiology can add to information overload	Conduct ongoing evaluation of existing and new sources and methods is necessary (ex participatory surv)
Infodemiology can provide feedback from public (perception)	Infodemiology complements traditional	Incorporate ways to rapidly identify false or misleading content
Infodemiology sources can be used for broader purposes than disease detection and surveillance	Focus on adding infodemiology for diseases where source has high potential to be valuable	Develop methods for threshold setting for infodemiology sources
The public has perceptions about these data types, so organizations need to transparent	Recognize challenges and limitations of data sources	Incorporate validity and reliability of sources in any report or tool

Importantly, training and education is critical for all three stakeholder groups who are going to be incorporating infodemiology into situational awareness. There is a need for a new profession or discipline within public health or adaption of existing public health professional training where these new data sources will be part of the curriculum. Ideally, the findings of this work should also be incorporated into continuing training for on-board public health professionals. Regardless of these new mechanisms, trained public health officials will need to be able to collaborate with providers, understand the context, assess event magnitude, evaluate credibility, and provide standardized interpretation (126). This adds another tool to the toolbox for public health staff within organizations of all levels and sizes. Additionally, many traditional disaster organizations have negative opinions regarding the accuracy, utility, and value of social media due to potential spread of misinformation (51)

and educating policy makers on the self-policing social media community might help reduce these concerns.

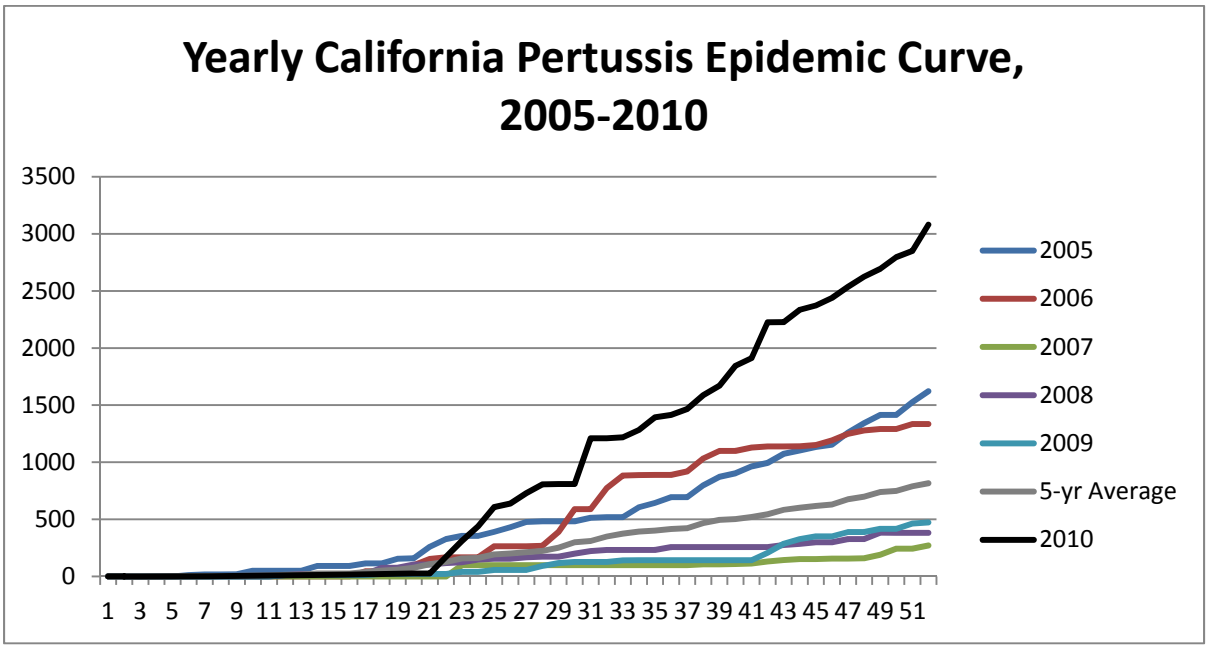
What Benefit Will This Bring

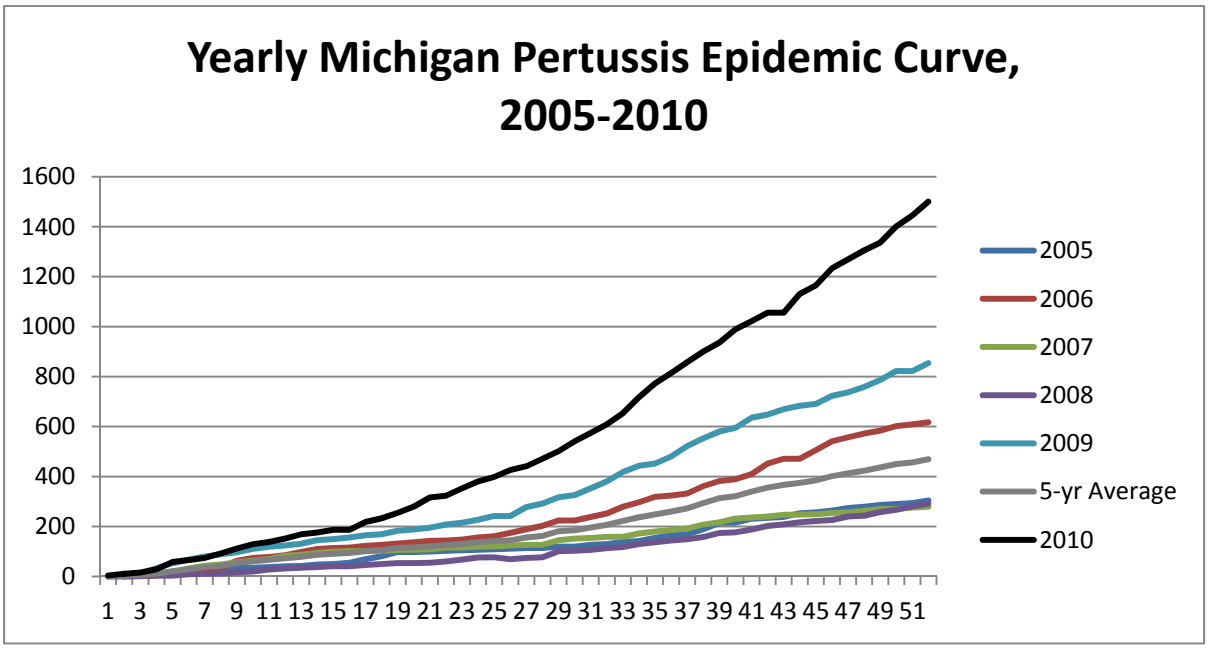
The development of a framework as described above will allow for incorporation of these, and other, novel data types. Utilizing these data sources effectively can result in improved timeliness and efficiency regarding detection of outbreaks. These sources can also provide characterization for disease outbreaks for which officials are aware but are still investigating. Any ability to know about outbreaks sooner and/or more completely can reduce morbidity and mortality by increasing the speed at which response actions occur. Knowing about an outbreak sooner is the only way to reduce the epidemics and pandemics of the future.

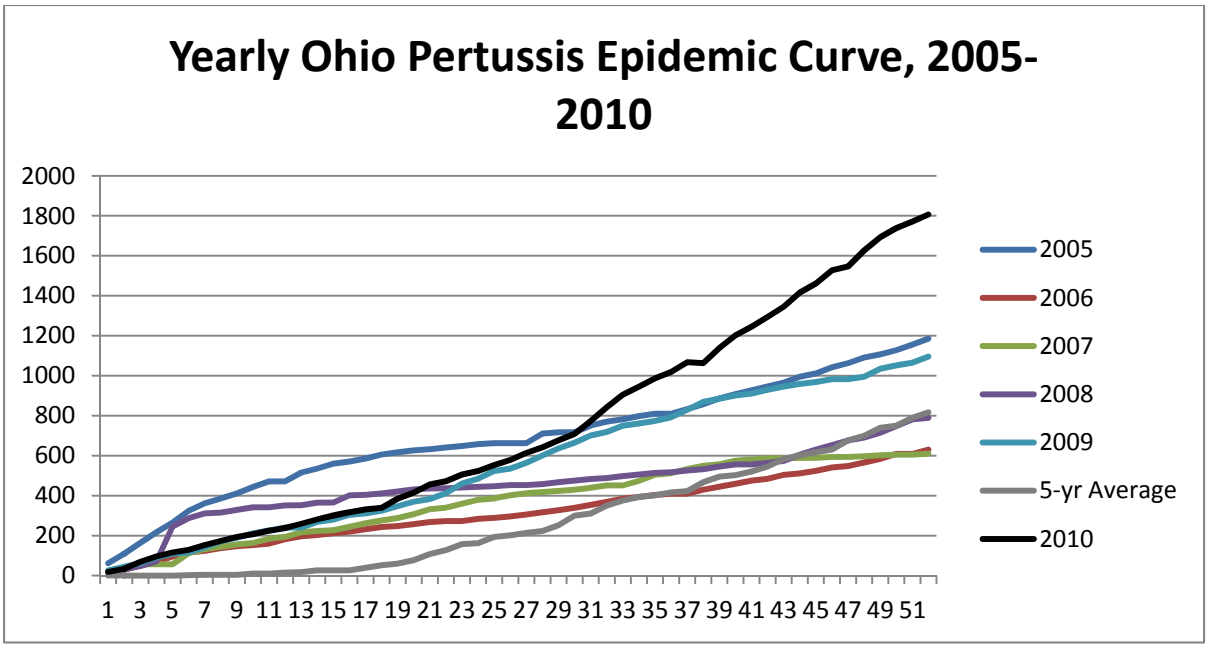
Appendix A

Pertussis 5-year Epidemiology Curves

The following curves show the pertussis epidemiology curves for each California, Michigan, and Ohio from 2005-2010. This research was focused on 2010, the black line, which is markedly different from each of the five previous years (2005-2009). The 2010 season was also more severe when compared to the average of the previous five years, as indicated by the gray line.






Yearly Ohio Pertussis Epidemic Curve, 2005-2010
2000
1800
1600
1400
1200
1000
800
600
400
200
0
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51
2005
2006
2007
2008
2009
5-yr Average
2010

Appendix B

List of Pertussis Terms

Within Google Insights*:

- Bordetella
- Pertussis
- "Whooping Cough"

*Google auto-corrects mis-spellings and associates close spellings with the actual term, so we do not need to accommodate for alternate (mis-) spellings in Google

Within Twitter

- "pertuss"
- "pertusi"
- "whoop"
- "whopp"
- "woop"
- "bordetell"

Appendix C

Survey Questions

#	Question	Intent	Responses	Theory
1	Which of the following have you ever used (select all that apply):	Determine if beliefs about various technologies are confounded by lack of knowledge or awareness of the tools	-Twitter -Facebook -Search Engines -Email News Alerts (like PubMed)	If someone has not used a tool, it is hard for him/her to assess its' value.
2	*[For any of the systems indicated above]:* Within the past month, how often have you used _____?	Determine if beliefs about various technologies are confounded by lack of knowledge or familiarity of the tools	*a) None* *b) 1-2 times* *c) 3-5 times* *d) 6-10 times* *e) More than 10 times*	If someone has not been exposed to a tool, it is hard for him/her to assess its' value.
3	*Select your level of agreement with the following statement:* Community members can provide valuable information about disease outbreaks in their community.	Regardless of the tools, how do you feel about crowd sourcing?	*a) Strongly Agree* *b) Somewhat Agree* *c) Neither agree nor disagree* *d) Somewhat Disagree* *e) Strongly Disagree*	Ensure potential confounder of lack of believe in the source is not the cause
4	How valuable in identifying **early indications or signals** of emerging disease outbreaks is information from -social media (like Twitter or Facebook)? -news media? -internet search patterns?	Can any of these sources provide sentinel or early indication information?	*a) Highly valuable* *b) Moderately Valuable* *c) Minimally Valuable* *d) Not Valuable* *e) Unsure*	Level 1 SA: perception from situational awareness feedback loop (55) page 6
5	How valuable in strengthening your **comprehension** of	Can any of these sources provide validation or context for an ongoing	*a) Highly valuable* *b)*	Level 2 SA: comprehension from situational

	disease outbreaks is information from -social media (like Twitter or Facebook)? -news media? -internet search patterns?	event?	*Moderately Valuable* *c) Minimally Valuable* *d) Not Valuable* *e) Unsure*	awareness feedback loop (55) p6
6	How valuable for informing **future projections** about disease outbreaks is information from -social media (like Twitter or Facebook)? -news media? -internet search patterns?	Can any of these sources enable predictions for an ongoing event?	*a) Highly valuable* *b) Moderately Valuable* *c) Minimally Valuable* *d) Not Valuable* *e) Unsure*	Level 3 SA: projection from situational awareness feedback loop (55) p6
7	In your experience, how often is confirmatory data necessary for making decisions during a disease outbreak?	Will you only utilize structured/official/traditional sources?	*a) Always* *b) Usually* *c) Rarely* *d) Never* *e) Unsure*	Do you act on your gut (57) or do you wait for evidence?
8	*Select your level of agreement with the following:* It is possible to get too much information to make a decision during a disease outbreak.	Are you a sponge for content and will attempt to gather any information?	*a) Strongly Agree* *b) Agree* *c) Neither agree nor disagree* *d) Disagree* *e) Strongly Disagree*	Do you believe there is a beneficial degree of ignorance (57)?
9	*Select your level of agreement with the following:* I am always looking for new types of data and information sources to inform public health response actions.	Are you broadly opposed to more content, regardless of the source?	*a) Strongly Agree* *b) Agree* *c) Neither agree nor disagree* *d) Disagree* *e) Strongly Disagree*	Opposite of #6 (59)
10	Please indicate your current work setting.		*a) State health department*	

			b) County or city health department	
11	Please indicate your age group.	Determine if age/generation is a confounder for beliefs and utilization about new technology	*a) Under 30* *b) 30-39* *c) 40-49* *d) 50-59* *e) 60 and over*	
12	What questions need to be answered or research needs to be completed for you to have insights that are more complete about novel data sources?			
13	What is critical for you to know or understand before you use a new information source to make decisions during a disaster?			

Appendix D

Definition of Terms

Crowdsourcing – “the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people and especially from the online community rather than from traditional employees or suppliers” (127)

Epidemic Intelligence – the use of both official sources (such as public health surveillance systems) and informal sources (such as electronic media and web-based tools) for the purpose of early warning and initial risk assessment (128)

Infodemiology – “the science of distribution and determinants of information in an electronic medium, specifically the Internet, with the ultimate aim to inform public health and public policy” (1)

Leading Indicator – a variable that is consistently shown to provide an early indication of an event; in this context the event is a disease outbreak (129)

Multistream Surveillance – “approach that monitors multiple sources of information and may also integrate them into a unified analytical framework” (9)

News Media – newspapers and magazines collectively

Official Reporting –any product produced by a government entity (local, state, federal, or international) including items marked “For Official Use Only”

Public Health – “the practice of preventing disease and promoting good health within groups of people, from small communities to entire countries” (130)

Quality – assessed as a measure of timeliness, relevance, validity, accuracy, variability, volume, geography, sensitivity, and specificity (68)

Situational Awareness – a person or organization's "ability to identify, process, and comprehend the critical information about an incident" (131); maintaining situational awareness requires continuous monitoring of information sources about actual incidents and developing or emerging hazards (131).

Social Media – "forms of electronic communication (as Web sites for social networking and microblogging) through which users create online communities to share information, ideas, personal messages, and other content (as videos) (132); examples include blogs, social networking, and news sharing sites.

Syndromic Surveillance – an approach to monitoring population health using pre-diagnostic data to identify unusual events that warrant further public health investigation (133)

Web 2.0 – the new culture of how people interact with the Web, relies on users to supply the content that other users demand (32); term for futuristic, emerging web-based capabilities

References

1. Eysenbach G. Infodemiology and infoveillance tracking online health information and cyberbehavior for public health. Am J Prev Med. 2011 May;40(5 Suppl 2):S154-8.

2. Buckeridge DL. Outbreak detection through automated surveillance: A review of the determinants of detection. J Biomed Inform. 2007 Aug;40(4):370-9.

3. Chin J, editor. Control of communicable diseases manual. 17th ed. Washington, DC: American Public Health Association; 2000.

4. Centers for Disease Control and Prevention (CDC). Summary of notifiable diseases--united states, 2010. MMWR Morb Mortal Wkly Rep. 2012 Jun 1;59(53):1-111.

5. FDA approves boostrix to prevent tetanus, diphtheria, and pertussis in older people [Internet]. Silver Spring, MD: FDA; 2011 [updated July 8, 2011; cited June 15, 2012]. Available from: http://www.fda.gov/NewsEvents/Newsroom/PressAnnouncements/ucm262390.htm.

6. Atkinson W, Wolfe S, Hamborsky J, editors. Epidemiology and prevention of vaccine-preventable diseases. 12th edition, 2nd printing ed. Washington, DC: Public Health Foundation; 2012.

7. Adelman RC, Verbrugge LM. Death makes news: The social impact of disease on newspaper coverage. J Health Soc Behav. 2000 Sep;41(3):347-67.

8. Boak MB, M'ikanatha NM, Day RS, Harrison LH. Internet death notices as a novel source of mortality surveillance data. Am J Epidemiol. 2008 Mar 1;167(5):532-9.

9. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Surveillance sans frontieres: Internet-based emerging infectious disease intelligence and the HealthMap project. PLoS Med. 2008 Jul 8;5(7):e151.

10. Brownstein JS, Freifeld CC, Chan EH, Keller M, Sonricker AL, Mekaru SR, et al. Information technology and global surveillance of cases of 2009 H1N1 influenza. N Engl J Med. 2010 May 6;362(18):1731-5.

11. Doyle TJ, Glynn MK, Groseclose SL. Completeness of notifiable infectious disease reporting in the united states: An analytical literature review. Am J Epidemiol. 2002 May 1;155(9):866-74.

12. Centers for Disease Control and Prevention (CDC). Summary of notifiable diseases--united states, 2001. MMWR Morb Mortal Wkly Rep. 2003 May 2;50(53):i,xxiv, 1-108.

13. Jajosky RA, Groseclose SL. Evaluation of reporting timeliness of public health surveillance systems for infectious diseases. BMC Public Health. 2004 Jul 26;4:29.

14. Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. PLoS One. 2009;4(2):e4378.

15. Madoff LC, Woodall JP. The internet and the global monitoring of emerging diseases: Lessons from the first 10 years of ProMED-mail. Arch Med Res. 2005 Nov-Dec;36(6):724-30.

16. Freifeld CC, Chunara R, Mekaru SR, Chan EH, Kass-Hout T, Ayala Iacucci A, et al. Participatory epidemiology: Use of mobile phones for community-based health reporting. PLoS Med. 2010 Dec 7;7(12):e1000376.

17. Nelson NP, Brownstein JS, Hartley DM. Event-based biosurveillance of respiratory disease in mexico, 2007-2009: Connection to the 2009 influenza A(H1N1) pandemic? Euro Surveill. 2010 Jul 29;15(30):19626.

18. Heymann DL, Rodier GR, WHO Operational Support Team to the Global Outbreak Alert and Response Network. Hot spots in a wired world: WHO surveillance of emerging and re-emerging infectious diseases. Lancet Infect Dis. 2001 Dec;1(5):345-53.

19. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection--harnessing the web for public health surveillance. N Engl J Med. 2009 May 21;360(21):2153,5, 2157.

20. Carneiro HA, Mylonakis E. Google trends: A web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis. 2009 Nov 15;49(10):1557-64.

21. Chan EH, Sahai V, Conrad C, Brownstein JS. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance. PLoS Negl Trop Dis. 2011 May;5(5):e1206.

22. Boulos MNK, Sanfilippo AP, Corley CD, Wheeler S. Social web mining and exploitation for serious applications: Technosocial predictive analytics and related technologies for public health, environmental and national security surveillance. Comput Methods Programs Biomed. 2010 OCT;100(1):16-23.

23. Chew C, Eysenbach G. Pandemics in the age of twitter: Content analysis of tweets during the 2009 H1N1 outbreak. PLoS One. 2010 Nov 29;5(11):e14118.

24. Anderson AL, Leffler K. Bedbug infestations in the news: A picture of an emerging public health problem in the united states. J Environ Health. 2008 May;70(9):24,7, 52-3.

25. American Red Cross. Social media in disasters and emergencies. Washington, DC: American Red Cross; 2011 Summer, 2011.

26. Takahashi Y, Miyaki K, Nakayama T. Analysis of news of the japanese asbestos panic: A supposedly resolved issue that turned out to be a time bomb. J Public Health (Oxf). 2007 Mar;29(1):62-9.

27. About HealthMap [Internet]. Boston, MA: HealthMap; cited January 5, 2013]. Available from: http://healthmap.org/about/.

28. Freifeld CC, Mandl KD, Reis BY, Brownstein JS. HealthMap: Global infectious disease monitoring through automated classification and visualization of internet media reports. J Am Med Inform Assoc. 2008 Mar-Apr;15(2):150-7.

29. Brownstein JS, Freifeld CC. HealthMap: The development of automated real-time internet surveillance for epidemic intelligence. Euro Surveill. 2007 Nov 29;12(11):E071129.5.

30. Polgreen PM, Chen Y, Pennock DM, Nelson FD. Using internet searches for influenza surveillance. Clin Infect Dis. 2008 Dec 1;47(11):1443-8.

31. Reis BY, Brownstein JS. Measuring the impact of health policies using internet search patterns: The case of abortion. BMC Public Health. 2010 AUG 25;10:514.

32. Lee BK. Epidemiologic research and web 2.0--the user-driven web. Epidemiology. 2010 Nov;21(6):760-3.

33. Yang AC, Huang NE, Peng C, Tsai S. Do seasons have an influence on the incidence of depression? the use of an internet search engine query data as a proxy of human affect. Plos One. 2010 OCT 28;5(10):e13728.

34. Breyer BN, Eisenberg ML. Use of google in study of noninfectious medical conditions. Epidemiology. 2010 Jul;21(4):584-5.

35. Surowiecki J, 1967-. The wisdom of crowds. New York: Anchor Books; 2005.

36. comScore releases december 2011 U.S. search engine rankings [Internet].: comScore [updated January 11, 2012. Available from: http://www.comscore.com/Press_Events/Press_Releases/2012/1/comScore_Releases_December_2011_U.S._Search_Engine_Rankings.

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

38. Fox S. The social life of health information, 2011. Washington, DC: Pew Internet & American Life Project; 2011 May 12, 2011.

39. Scanfeld D, Scanfeld V, Larson EL. Dissemination of health information through social networks: Twitter and antibiotics. Am J Infect Control. 2010 Apr;38(3):182-8.

40. Fox S. Online health search 2006. Washington, DC: Pew Internet & American Life Project; 2006 October 29, 2006.

41. Valdivia A, Lopez-Alcalde J, Vicente M, Pichiule M, Ruiz M, Ordobas M. Monitoring influenza activity in europe with google flu trends: Comparison with the findings of sentinel physician networks - results for 2009-10. Eurosurveillance. 2010 JUL 22;15(29):2-7.

42. Jones JH. Early assessment of anxiety and behavioral response to novel swine-origin influenza A (H1N1). PloS one. 2009;4(12):e8032.

43. Paul MJ, Dredze M. You are what you tweet: Analyzing twitter for public health. Fifth international AAAI conference on weblogs and social media; July 17, 2011; Barecelona, Spain. AAAI; 2011.

44. Dredze M. How social media will change public health. Intelligent Systems, IEEE. 2012;27(4):81-4.

45. Heighington A. Homeland security in real-time: The power of the public and mobile technology. Homeland security affairs. 2011;7(13):1.

46. Gupta A, Kumaraguru P. Credibility ranking of tweets during high impact events. Proceedings of the 1st workshop on privacy and security in online social media; ACM; 2012.

47. Vieweg SE. Situational awareness in a mass emergency: A behavioral and linguistic analysis of microblogged communications [dissertation]. Colorado: University of Colorado; 2012.

48. Lindsay BR. Social media and disasters: Current uses, future options, and policy considerations. Washington, D.C.: Congressional Research Service; 2011 September 6, 2011. Report No.: 7-5700.

49. Poblete B, Garcia R, Mendoza M, Jaimes A. Do all birds tweet the same?: Characterizing twitter around the world. Proceedings of the 20th ACM international conference on information and knowledge management; ACM; 2011.

50. Kamel Boulos MN, Resch B, Crowley DN, Breslin JG, Sohn G, Burtner R, et al. Crowdsourcing, citizen sensing and sensor web technologies for public and environmental health surveillance and crisis management: Trends, OGC standards and application examples. Int J Health Geogr. 2011 Dec 21;10:67,072x-10-67.

51. Keim ME, Noji E. Emergent use of social media: A new age of opportunity for disaster resilience. Am J Disaster Med. 2011 Jan-Feb;6(1):47-54.

52. Twitter update 2011 [Internet]. Washington, DC: Pew Internet & American Life Project; 2011 [updated June 1, 2011; cited 2/12/12]. Available from: http://pewinternet.org/Reports/2011/Twitter-Update-2011/Main-Report.aspx.

53. Lenhart A, Fox S. Twitter and status updating. memo. http://pewinternet.org/: Pew Internet & American Life Project; 2009 February 12, 2009 4pm. Report No.: Pew Internet Project Data Memo.

54. Moore RJ. Twitter data analysis: An investor's perspective. TechCrunch. 2009 October 5, 2009.

55. Situation awareness : Analysis and measurement. edited by Mica R. Endsley,Daniel J.Garland., Endsley MR, Garland DJ, editors. Mahwah, N.J.: Lawrence Erlbaum; 2000.

56. Hawkins J, 1957-. On intelligence. Blakeslee S, editor. New York: Henry Holt and Company; 2005.

57. Gigerenzer G. Gut feelings : The intelligence of the unconscious. New York: Viking; 2007.

58. Lehrer J. How we decide. Boston: Houghton Mifflin Harcourt; 0003.

59. Gladwell M, 1963-. Blink : The power of thinking without thinking. New York: Little, Brown and Co; 2005.

60. Klein GA. Sources of power : How people make decisions. Cambridge, Mass.: MIT Press, 1999; 1998.

61. Kahneman D, Tversky A. On the psychology of prediction. Psychol Rev. 1973 07;80(4):237-51.

62. Groopman JE. How doctors think. Boston: Houghton Mifflin; 2007.

63. Nasrullah M, Muazzam S. Newspaper reports: A source of surveillance for burns among women in pakistan. Journal of Public Health. 2010 JUN;32(2):245-9.

64. Barss P, Subait OM, Ali MH, Grivna M. Drowning in a high-income developing country in the middle east: Newspapers as an essential resource for injury surveillance. J Sci Med Sport. 2009 Jan;12(1):164-70.

65. Wilson JM, Iannarone M, Wang C. Media reporting of the emergence of the 1968 influenza pandemic in hong kong: Implications for modern-day situational awareness. Disaster Med Public Health Prep. 2009 Dec;3 Suppl 2:S148-53.

66. Jensen JD, Moriarty CM, Hurley RJ, Stryker JE. Making sense of cancer news coverage trends: A comparison of three comprehensive content analyses. J Health Commun. 2010 Mar;15(2):136-51.

67. Chan EH, Brewer TF, Madoff LC, Pollack MP, Sonricker AL, Keller M, et al. Global capacity for emerging infectious disease detection. Proc Natl Acad Sci U S A. 2010 Nov 29.

68. Brownstein JS, Freifeld CC, Reis BY, Mandl KD. Evaluation of online media reports for global infectious disease intelligence. Adv Disease Surv. 2007;4(3).

69. Reilly AR, Iarocci EA, Jung CM, Hartley DM, Nelson NP. Indications and warning of pandemic influenza compared to seasonal influenza. Adv Disease Surv. 2008;5(190).

70. Connor SM, Wesolowski K. Newspaper framing of fatal motor vehicle crashes in four midwestern cities in the united states, 1999-2000. Inj Prev. 2004 Jun;10(3):149-53.

71. Barss P, Subait OM, Al-Ali M, Grivna M. Drowning in a high-income developing country in the middle east: Newspapers as an essential resource for injury surveillance. Journal of Science and Medicine in Sport. 2009;12(1):164-70.

72. Collier N. What's unusual in online disease outbreak news? J Biomed Semantics. 2010 Mar 31;1(1):2.

73. Ghaffar A, Hyder AA, Bishai D. Newspaper reports as a source for injury data in developing countries. Health Policy Plan. 2001 Sep;16(3):322-5.

74. Habel MA, Liddon N, Stryker JE. The HPV vaccine: A content analysis of online news stories. J Womens Health (Larchmt). 2009 Mar;18(3):401-7.

75. La Bash HA, Vogt DS, King LA, King DW. Deployment stressors of the iraq war: Insights from the mainstream media. J Interpers Violence. 2009 Feb;24(2):231-58.

76. Cooper CP, Mallon KP, Leadbetter S, Pollack LA, Peipins LA. Cancer internet search activity on a major search engine, united states 2001-2003. J Med Internet Res. 2005 Jul 1;7(3):e36.

77. Leffler CT, Davenport B, Chan D. Frequency and seasonal variation of ophthalmology-related internet searches. Can J Ophthalmol. 2010 Jun;45(3):274-9.

78. Seifter A, Schwarzwalder A, Geis K, Aucott J. The utility of "google trends" for epidemiological research: Lyme disease as an example. Geospatial Health. 2010 MAY;4(2):135-7.

79. Pelat C, Turbelin C, Bar-Hen A, Flahault A, Valleron A. More diseases tracked by using google trends. Emerg Infect Dis. 2009 Aug;15(8):1327-8.

80. Boyle JR, Sparks RS, Keijzers GB, Crilly JL, Lind JF, Ryan LM. Prediction and surveillance of influenza epidemics. Med J Aust. 2011 Feb 21;194(4):S28-33.

81. Valdivia A, Monge-Corella S. Diseases tracked by using google trends, spain. Emerg Infect Dis. 2010 Jan;16(1):168.

82. Wilson N, Mason K, Tobias M, Peacey M, Huang QS, Baker M. Interpreting google flu trends data for pandemic H1N1 influenza: The new zealand experience. Euro Surveill. 2009 Nov 5;14(44):19386.

83. Wilson K, Brownstein JS. Early detection of disease outbreaks using the internet. CMAJ. 2009 Apr 14;180(8):829-31.

84. Bentley RA, Ormerod P. A rapid method for assessing social versus independent interest in health issues: A case study of 'bird flu' and 'swine flu'. Soc Sci Med. 2010 Aug;71(3):482-5.

85. McCarthy MJ. Internet monitoring of suicide risk in the population. J Affect Disord. 2010 May;122(3):277-9.

86. Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance. PLos One. 2009:e4378-.

87. Keelan J, Pavri V, Balakrishnan R, Wilson K. An analysis of the human papilloma virus vaccine debate on MySpace blogs. Vaccine. 2010 Feb 10;28(6):1535-40.

88. World Health Organization. Weekly epidemiological record. Geneva, Switzerland: World Health Organization; 2000 7 January 2000. Report No.: 1.

89. Wilson JM 5, Polyak MG, Blake JW, Collmann J. A heuristic indication and warning staging model for detection and assessment of biological events. J Am Med Inform Assoc. 2008 Mar-Apr;15(2):158-71.

90. Benson R, Blach-Ørsten M, Powers M, Willig I, Zambrano SV. Media systems online and off: Comparing the form of news in the united states, denmark, and france. J Commun. 2012;62(1):21-38.

91. MMWR tables (NNDSS**) [Internet]. Atlanta, GA: Centers for Disease Control and Prevention; cited 8/19/12]. Available from: http://wonder.cdc.gov/mmwr/mmwrmorb2.asp?mmwr_year=2010&mmwr_week=35.

92. Centers for Disease Control and Prevention. Provisional national notifiable diseases surveillance. Atlanta, GA: Centers for Disease Control and Prevention; 2010 September 27, 2010.

93. Mekaru S. Request for HealthMap info. 2012 August 15, 2012;Email.

94. [Internet].: Google; 2011. Available from: http://www.google.ca/intl/en/trends/about.html.

95. Google insights for search beta [Internet]. California: Google; 2012; cited August 20, 2012]. Available from: http://www.google.com/insights/search/.

96. Google insights: What do the numbers on the graph mean? [Internet]. California: Google; 2012; cited August 20, 2012]. Available from: http://support.google.com/insights/bin/answer.py?hl=en-US&answer=87285.

97. Google insights for search: Is the data normalized [Internet]. California: Google; 2012; cited August 20, 2012]. Available from: http://support.google.com/insights/bin/bin/answer.py?hl=en&answer=87284.

98. About us [Internet]. Boulder, CO: Gnip, Inc; 2013; cited January 27, 2013]. Available from: http://gnip.com/about/.

99. de Leeuw ED, Hox JJ, Dillman DA, European Association of Methodology. International handbook of survey methodology. New York: L. Erlbaum Associates; 2008.

100. About ISDS [Internet]. Brighton, MA: International Society for Disease Surveillance; cited January 27, 2013]. Available from: http://www.syndromic.org/about-isds.

101. SAS Institute Inc. SAS for windows. 2010;9.3 TS Level 1M1.

102. Johnson HA, Wagner MM, Hogan WR, Chapman W, Olszewski RT, Dowling J, et al. Analysis of web access logs for surveillance of influenza. Stud Health Technol Inform. 2004;107(Pt 2):1202-6.

103. Hirsch RP, Poneleit KM. Biostatistical applications in health research. Stat-Aid, Inc; 2008.

104. Keller M, Blench M, Tolentino H, Freifeld CC, Mandl KD, Mawudeku A, et al. Use of unstructured event-based reports for global infectious disease surveillance. Emerg Infect Dis. 2009 May;15(5):689-95.

105. Wu H. Systemic determinants of international news coverage: A comparison of 38 countries. J Commun. 2000 -06-01;50(2):110-30.

106. Bennett WL. News : The politics of illusion. Boston: Longman; 2012.

107. Walgreens offering pertussis (whooping cough) vaccinations daily at more than 150 stores throughout california [Internet]. Deerfield, IL: Walgreens.com; 2010 [updated September 17, 2010; cited January 1, 2013]. Available from: http://news.walgreens.com/article_print.cfm?article_id=5339.

108. Twitter statistics for 2010: An in-depth report at Twitter´s growth 2010, compared with 2009 [Internet]. Ontario, Canada: Sysomos; 2010 [updated December 2010; cited February 26, 2012]. Available from: http://www.sysomos.com/insidetwitter/twitter-stats-2010/.

109. Kwak H, Lee C, Park H, Moon S. What is twitter, a social network or a news media? Proceedings of the 19th international conference on world wide web; ACM; 2010.

110. Quenel P, Dab W, Hannoun C, Cohen JM. Sensitivity, specificity and predictive values of health service based indicators for the surveillance of influenza A epidemics. Int J Epidemiol. 1994 Aug;23(4):849-55.

111. Shih T, Fan X. Comparing response rates from web and mail surveys: A meta-analysis. Field Methods. 2008;20(3):249-71.

112. United States Congress Senate Committee,on Health. Pandemic and all-hazards preparedness act : Report (to accompany S. 3678). Washington, D.C.: U.S. G.P.O; 2006.

113. The White House. National strategy for biosurveillance. Washington, DC: 2012 July 31, 2012.

114. Centers for Disease Control Office of Preparedness and Response Biosurveillance Coordination Unit. National biosurveillance strategy for human health. Atlanta, GA: US Department of Health & Human Services; 2010 February 2010. Report No.: version 2.0.

115. Crowe A. The social media manifesto: A comprehensive review of the impact of social media on emergency management. Journal of Business Continuity & Emergency Planning. 2011 02;5(1):409-20.

116. Burke DS, Cauchemez S, Cummings DAT, Ferguson NM, Fraser C, Iamsirithaworn S, et al. Strategies for containing an emerging influenza pandemic in southeast asia. Nature. 2005 09/08; 2012/1;437:209.

117. Sutton J, Palen L, Shklovski I. Backchannels on the front lines: Emergent uses of social media in the 2007 southern california wildfires. Proceedings of the 5th international ISCRAM conference; Washington, DC; 2008.

118. Meier P. Debating the value of tweets for disaster response (intelligently). iRevolution. 2012 December 17, 2012.

119. Funk S, Gilad E, Watkins C, Jansen VA. The spread of awareness and its impact on epidemic outbreaks. Proc Natl Acad Sci U S A. 2009 Apr 21;106(16):6872-7.

120. Couzin ID, Franks NR, Krause J, Levin SA. Effective leadership and decision-making in animal groups on the move. Nature. 2005 02/03; 2012/1;433:513+.

121. Wagner MM, Tsui FC, Espino JU, Dato VM, Sittig DF, Caruana RA, et al. The emerging science of very early detection of disease outbreaks. J Public Health Manag Pract. 2001 Nov;7(6):51-9.

122. Prier K, Smith M, Giraud-Carrier C, Hanson C. Identifying health-related topics on twitter: An exploration of tobacco-related tweets as a test topic. International conference on social computing, behavioral-cultural modeling, & prediction (SBP11); March 2011; ; 2011.

123. Quercia D, Askham H, Crowcroft J. TweetLDA: Supervised topic classification and link prediction in twitter. Proceedings of the 4th ACM international conference on web science (WebSci); ; 2012.

124. Kirkpatrick M. Google CEO schmidt: "people aren't ready for the technology revolution". 2010 August 4, 2010.

125. Paul MJ. A model for mining public health topics from twitter. Health (London, England : 1997). 2011;11:16.

126. Khan AS, Fleischauer A, Casani J, Groseclose SL. The next public health revolution: Public health information fusion and social networks. Am J Public Health. 2010 Jul;100(7):1237-42.

127. Crowdsourcing [Internet].: Merriam Webster; 2012; cited 03/04/2012]. Available from: http://www.merriam-webster.com/dictionary/crowdsourcing.

128. Linge JP, Steinberger R, Weber TP, Yangarber R, van der Goot E, Al Khudhairy DH, et al. Internet surveillance systems for early alerting of health threats. Euro Surveill. 2009 Apr 2;14(13):19162.

129. Bloom RM, Buckeridge DL, Cheng KE. Finding leading indicators for disease outbreaks: Filtering, cross-correlation, and caveats. J Am Med Inform Assoc. 2007 Jan-Feb;14(1):76-85.

130. Get the facts: What is public health? Washington, DC: American Public Health Association; 2011.

131. National response framework. Washington, DC: US Department of Homeland Security; 2008 January 2008.

132. Social media definition and more from the free merriam-webster dictionary [Internet]. Springfield, MA: Merriam-Webster, Incorporated; 2012; cited December 7, 2012]. Available from: http://www.merriam-webster.com/dictionary/social%20media.

133. Elbert Y, Burkom HS. Development and evaluation of a data-adaptive alerting algorithm for univariate temporal biosurveillance data. Stat Med. 2009 Nov 20;28(26):3226-48.