Role of Social Media in Early Warning of Norovirus Outbreaks: A Longitudinal Twitter-Based Infoveillance

Ahmed H. YoussefAgha, Wasantha P. Jayawardene, David K. Lohrmann

Abstract: The purpose of this study was to determine the trend in daily norovirus-related keyword utilization on twitter and to develop an experimental computational model that can accurately predict outbreaks in real-time. Data were collected from twitter within an accessible limit (1%) between February 1 and May 5, 2012 using seven keywords. Data were analyzed to determine the trend of daily norovirus-related keywords utilization on twitter on daily bases. Because of the trend lines on time were expected to be non-linear, a polynomial of degree five was used to model the trends in the norovirus hashtag separately by week. We also explored the correlation between norovirus hashtag utilization on twitter and other related hashtags. For categorical data analysis, each hashtag distribution was transformed into a binomial distribution. Nonparametric test of Wilcoxon Scores (Rank Sums) was used to compare norovirus days with different codes. Chi-Square test was used to explore associations between norovirus and other hashtags. Probability of the "norovirus" hashtags occurring above the daily mean on a day with "fever" hashtags above the daily mean was 0.467 (p=0.0433), whereas that for "outbreak" was 0.625 (p=0.027). "Norovirus" hashtag had the highest correlation with "fever" hashtag, followed by "outbreak", "throwing up", and "sick" hashtags. A statistically significant difference between "fever" and "sick" keywords was found in relation to utilization of the "norovirus" hashtag. A non-linear regression equation, using a polynomial of degree six, was formed for each of the four short term extrapolation periods.

Keywords: Norovirus; Outbreak; Twitter; Hashtags

I. INTRODUCTION

Acute gastroenteritis, usually accompanied by diarrhea, nausea, vomiting abdominal pain, and/or fever, is one of the leading causes of morbidity in the United States. Approximately 179 million cases of acute gastroenteritis leading to approximately 0.6 million hospital admissions and 5,000 deaths occur every year [1]. As detection of viruses, unlike bacteria, in foods is very difficult, the best way of identifying the causative agent in the majority of outbreaks is the epidemiological analysis of patients [2].

Transmission of noroviruses occurs mainly through the fecal-oral route, either directly from person to person [3] or indirectly through contaminated food [4] water [5], surfaces [6], or animals [7]. Airborne transmission of infectious droplets can also occur during vomiting [8]. Most norovirus outbreaks occur in locations with high density of susceptible individuals is high, such as hospitals [9], elderly homes [10], and military bases [11], as well as in settings where turnover of vulnerable individuals, such as hotels [12], restaurants [13], and cruise ships [14].

People with norovirus are contagious for three days from the onset of symptoms to, although contagiousness may persist for up to two weeks after recovery from symptoms. Major symptoms are vomiting (more common among children) and diarrhea (more common among adults) several times a day [15]. Nausea, abdominal cramps, headache, fever, chills, and myalgia may also present as associated symptoms [15]. Winter vomiting disease, a condition characterized with vomiting alone, can also occur [16]. Due to the nature of symptoms, people usually call norovirus infection "stomach flu" [14, 17] or sometimes "gastric flu" [12]. Most people recover from symptoms within 12-60 hours, although dehydration can be problematic among young children, the elderly, and people with debilitating illnesses [15]. Norovirus is not a nationally notifiable disease in the US, because testing for the disease is not generally available in hospitals and doctor's offices. Therefore, norovirus is usually diagnosed only when an outbreak of symptoms is reported to CDC [18]. A norovirus outbreak is defined as the occurrence of two or more similar cases that are linked epidemiologically; for example, ingestion of a common food [18].

Traditionally, newspapers, radio, and television are the major sources of information from public health agencies to the public and play a large role in risk communication during outbreaks [19]. However, internet was the most frequently used source of information about the H1N1 pandemic in 2009 [20]. The type of disease surveillance that utilizes online contents is called infoveillance [21]. Because twitter has short text status updates with <140 characters (tweets) that users share with followers, it's a candidate for longitudinal infoveillance text mining [19]. Longitudinal mining of tweets allows identification of changes in public responses [19], as well as early warning and detection of outbreaks, such as swine flu [22].

Ahmed H. YoussefAgha is with Department of Epidemiology and Biostatistics, School of Public Health Bloomington, Indiana University; Address: SPHB C108, 1025 E 7th Street, Bloomington, IN 47405, USA; Tel: 1-812-369-9798; E-mail: ahmyouss@indiana.edu

Wasantha P. Jayawardene (corresponding author) is with Department of Applied Health Science, School of Public Health Bloomington, Indiana University; Address: SPHB C116, 1025 E 7th Street, Bloomington, IN 47405, USA; Tel: 1-812-272-9136; E-mail: wajayawa@indiana.edu

David K. Lohrmann is with Department of Applied Health Science, School of Public Health Bloomington, Indiana University; Address: SPHB C116, 1025 E 7th Street, Bloomington, IN 47405, USA; Tel: 1-812-856-5101; E-mail: dlohrman@indiana.edu

II. METHODS

A. Purpose of the Study

The main purpose of our study is to investigate the trend of norovirus-related keyword utilization via twitter on daily basis and to develop an experimental computational model that can accurately predict outbreaks in real-time. A norovirus outbreak was identified and tracked through longitudinal mining and analysis of twitter data. Based on findings from previously published studies, tweets between 02/01/2012 and 05/02/2012 were archived for analysis.

B. Methodology

Our intent was to develop an infectious diseases monitoring system comprised of four dimensions: (1) a tweet classifier, which instantly monitored and analyzed an incoming tweet to determine whether it was disease-related; (2) a disease classifier, which extracted all disease features from each relevant tweet and identify which disease is being tracked; (3) a named entity recognition analyzer, which was responsible for extracting text available on web-sites that are referred to by twitter users; and (4) a data mining and alert generator - a software component that will generate disease alerts when necessary along with periodical reports.

The tweet classifier was responsible for performing two tasks (1) capturing live tweets in real-time and (2) analyzing captured tweets to determine whether a tweet was relevant to the scope of this project. Tweet classifier was trained by analyzing manually selected tweets using a machine learning algorithms to be discussed later. As result of the classification process, irrelevant tweets were ignored and relevant tweets were filtered in and stored for further multi-category classification.

The disease classifier processed the relevant tweets and mapped them into their most related disease. Therefore, each tweet was analyzed to extract disease features. However, to be able to classify a tweet as being related to one disease or another we developed a manually large training set that had sufficient data about the symptoms of the diseases we were monitoring. For each disease of interest, a profile was developed using tweet features and literature features.

The named entity recognition analyzer was responsible for extracting text available on web-sites referred to by Twitter users. Though not all tweets will have an embedded URL, we took advantage of this available detailed information when it was provided. News articles, for example, revealed the name of a disease that was tracked. It was very important also to identify the other entities (organism, person, percentage, quantity, and location) mentioned in a news article and then link them to the literature, the host organism, and the place of occurrence.

The data mining and alert generator was the most important component for our system. After tweets were classified and weather data was tracked for the subject region of the tweets, a series of data mining events executed to predict whether the disease was going to spread. The data mining and alert generator predicted the magnitude of spread given the location of the disease and rate of spread. We used statistical data mining methods and techniques to accomplish this task and accessed extensive computational resources to derive a complex model that enabled analysis of the resources needed to generate appropriate alerts.

C. Keywords

The "#" symbol in twitter is called a hashtag, used to mark keywords or topics in a tweet. It is created organically by twitter users as a way of classifying messages. According to the literature, "diarrhea", "throwing up", "nausea", "stomach pain", "fever", "headache", and "body ache" were chosen as hashtags for this twitter study. Additional hashtags, with similar meanings, i.e., "stomach flu", "sick", "throw up", and "outbreak", were also included.

D. Data Collection

We subscribed to the services available at Indiana University Pervasive Technology Institute for our system as a high performance application. Data were collected from twitter within an accessible limit (1%) between February 1, 2012 and May 5, 2012 using the keywords mentioned above. A sample size of 27 days was determined as the number needed to study the disease trend in four weeks. Therefore, a period of four weeks between February 1st and May 5th was randomly selected. Twitter messages sent within four week time frame were subjected to investigation.

E. Analysis

Data were analyzed to determine the trend of norovirus-related keywords utilization via twitter on a daily bases. Because the trend lines on time were expected to be non-linear, polynomial of degree five was used to model the trends of the norovirus hashtag separately for each week. So, for non-linear short term extrapolation and/or interpolation, each polynomial required 6-7 points (i.e., 6-7 days) to be developed. The non-linear polynomial of degree six can be used for short term extrapolation. The ability to access greater amounts of data (currently limited to 1% of all twitter) would have enhanced the extrapolation.

First, four conditional probabilities were evaluated:
a) pr (dayswithnorovirus>mean|dayswithfever>mean)
b) pr (dayswithnorovirus>mean|dayswithsick>mean)
c) pr (dayswithnorovirus>mean|dayswithvomiting>mean)
d) pr (dayswithnorovirus>mean|dayswithoutbreak>mean)

Then, the correlation between norovirus hashtag utilization on twitter with the other related hashtags was explored. For categorical data analysis, each hashtag distribution was transformed into a binomial distribution (table-1). For each selected study day, if a keyword hashtag frequency was less than or equal to the mean value of the same keyword, it was coded 1, and if the frequency was greater than the mean value, it was coded 2. Nonparametric test of Wilcoxon Scores (Rank Sums) was used to compare norovirus days with code 2 to norovirus days with code 1. Chi-Square test was used to assess associations between norovirus days (coded 1 or 2) and other hashtag (coded 1 or 2) such as fever, sick, outbreak, and vomiting.

III. RESULTS

A mean frequency of over 1700 hashtags (table-2) of “throwing-up”, “sick”, “headache”, and “fever” per day were found. Daily mean of other hashtags were between 14 and 27, except “diarrhea”, which was reported only 32 times during the 27-day period. The probability that more than 15 norovirus hashtags (daily mean for “norovirus”) occurred on a day with “fever” hashtag frequency exceeded 155 (daily mean for “fever”) was 0.467 and was statistically significant (p=0.0433). Similarly, the probability that more than 15 “norovirus” hashtags occurred on a day with “outbreak” hashtag frequency exceeded 23 (daily mean for “outbreak”) was 0.625 (p=0.027). However, none of the remaining norovirus-related hashtags (diarrhea, sick, headache, throwing-up, vomiting) had a statistically significant association with norovirus hashtags.

“Norovirus” hashtag had a moderate correlation with the seven other related hashtags collected during the four-week period (table-3). “Norovirus” hashtag had the highest correlation with “fever” hashtag (r=0.396), followed by correlations with “outbreak” (r=0.374), “throwing up” (r=0.281), “sick” (r=0.277), “headache” (r=0.258), “diarrhea” (r=0.180), and “vomiting” (r=0.155) hashtags.

Nonparametric test Wilcoxon Scores found that each of the “fever” and “sick” keywords significantly differed (p<0.05) for the days grouped by code 1 and the days grouped by code 2 in relation to utilization of “norovirus” hashtag. If at least two conditions of “fever”>155, “sick”>5395, and “vomiting”>76 were satisfied, and at the same time, if “outbreak”≤23, then the probability of “norovirus” hashtag being greater than 15 was 77.8%.

TABLE 1
TRANSFORMATION OF HASHTAG FREQUENCY INTO A BINARY DISTRIBUTION

Day*	Norovirus (μ=15)	Diarrhea (μ =1.4)	Fever (μ =155)	Sick (μ=5395)	Headache (μ =1718)	Throwing up (μ = 5420)	Vomiting (μ =26)	Outbreak (μ =23)
1st Feb	1	1	1	1	1	1	1	1
2nd Feb	1	2	1	1	1	1	1	1
3rd Feb	1	1	2	2	2	1	2	2
:	:	:	:	:	:	:	:	:
10th Feb	2	2	2	2	2	2	1	2
11th Feb	2	1	1	2	2	2	1	1
:	:	:	:	:	:	:	:	:

* Only a sample of the 27-day study period is shown in the table

TABLE 2
MEAN AND STANDARD DEVIATION OF NOROVIRUS AND OTHER RELATED HASHTAGS PER DAY

	Norovirus	Diarrhea	Fever	Sick	Outbreak	Headache	Throwing - up	Vomiting
Total	405	32	4194	145669	618	46379	145852	704
Mean	15	1.2	155	5395	23	1718	5402	26
SD	22	1	68	2137	27	603	1994	11

TABLE 3
CORRELATIONS AMONG NOROVIRUS AND OTHER RELATED HASHTAGS

	Norovirus	Diarrhea	Fever	Sick	Outbreak	Headache	ThrowingUp
Diarrhea	0.180	1					
Fever	0.396	0.309	1				
Sick	0.277	0.301	0.930	1			
Outbreak	0.374	0.000	0.472	0.246	1		
Headache	0.258	0.312	0.890	0.943	0.349	1	
ThrowingUp	0.281	0.214	0.775	0.841	0.413	0.927	1
Vomiting	0.155	0.247	0.695	0.762	0.319	0.812	0.832

Four periods of surveillance using the hashtags between February 1, 2012 and May 5, 2012: February 1–8, February 9–14, February 29–March 5, April 26–May 2 (figure 1). A non-linear regression equation was formed for each of the four periods for short term extrapolation; long term extrapolation was impossible as the trend was non-linear. A non-linear polynomial of degree six could be used for short term extrapolation. If the study could be expanded to access more than 1% of all twitter, the extrapolation would be enhanced.

IV. DISCUSSION

When twitter-based systems first appeared in epidemic surveillance, they were criticized for the possibility of producing exaggerated or misleading reports, lack of specificity (false positives), and extreme sensitivity to external forces such as unpredictable media interests. These condemnations are still valid, although they are recognized and adjusted to the extent possible. Despite these limitations, twitter-based epidemic surveillance is an irreplaceable resource for early warning of emerging outbreaks because of its sensitivity, availability, convenience, and transparency. Therefore, twitter has become a useful tool in the worldwide epidemic surveillance system. Moreover, findings of this study are highly compatible with the norovirus-related keyword search in “Google Trends” during the first half of 2012.

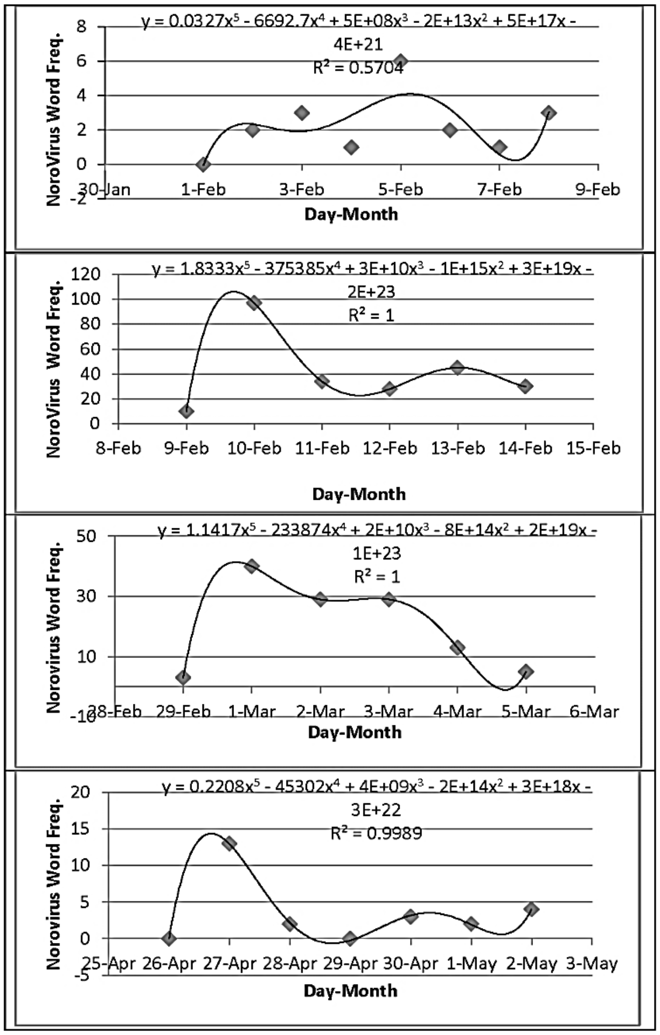

Figure 1: Short Term Extrapolation with Non-Linear Regression Equations for Each of the Four Periods between February 1, 2012 and May 2, 2012.

Several tweets-related challenges exist. High volumes of tweets are generated by millions of users tweeting in real-time simultaneously. This in turn required a high computational power to store and process incoming tweets. At its core, our model relied on the classification algorithms of tweets so that any tweet could be identified as relevant and, when relevant, could be further classified as to what disease was tracked. Finally, multi-lingual tweets are a challenge, because tweet users don't necessary share their status updates in English. Millions of twitter users share information in various languages, such as Spanish, French, Arabic, etc. More importantly, many users used less formal languages, acronyms, and even short hand. As a first step, we only focused on the tweets that were in English.

The geographic location, timing, and size of each norovirus outbreak may vary, complicating efforts to produce reliable and timely estimates of norovirus activity using traditional time series models. Epidemics are difficult to anticipate. Using actual tweet contents, which often reflected the user's perceived discomfort, when they were tweeting about their symptoms, we devised an estimation method based on well-understood statistical methods. The accuracy of the resulting real-time norovirus outbreak forecasting demonstrated that the subset of tweets identified and used in the models applied in the current study contained data associated with norovirus activity.

The current twitter-based model attempted to forecast norovirus activity. Because results generated by the current study could be available to public health officials as soon as the data are captured online, the forecast is potentially available at a much earlier time than ordinary public health alerts. Although it is possible to gather epidemic data in real time from hospital visits, drug purchase at pharmacies, and from school absenteeism, doing so at a national level would require combining data from different geographic areas and from multiple institutions/firms, a considerable data collecting burden. In contrast, twitter data are easily and efficiently collected and processed automatically in real time.

Despite these findings, this study has several limitations. First, the use of twitter is neither uniform across regions or time. Usually, Mondays are the busiest day for twitter traffic, while the lowest number of tweets is observed on Sundays. Large cities on east and west coasts produce far more tweets per person than cities in the Midwest or in other countries. In places where tweets are less frequent, the accuracy of our model may be low. Another limitation is that we only had 27 days of sampled data. Inclusion of more seasons, especially non-epidemic seasons, should help improve the accuracy of our norovirus estimates. Moreover, no comparable data, such as survey results, are available to validate our results. For example, absence of a detectable signal may indicate an apathetic public or a lack of knowledge. Therefore, we propose future studies to confirm our results with autocorrelated data.

The exact demographics of the twitter population are different from the general population. "Pew Research Center's Internet and American Life Project - Winter 2012 Tracking Survey, January 20 – February 19, 2012" (N=2,253), which coincided with the period of current study, showed that 15% of internet-users used twitter at some time and 8% of internet-users used twitter on a typical day. On a typical day, 20% of people in 18-24 year age group used twitter, with the percentage gradually decreasing with increasing age: 11% in 25-34 age group, 9% in 35-44 age group, 3% in 45-54 age group, 4% in 55-64 age group, and 1% in 65+ age group. According to the same survey, 15% of women and 14% of men used twitter, whereas 28% of non-Hispanic blacks, 14% of Hispanics, and 12% of non-Hispanic whites used it. People with no high school diploma had the highest twitter usage (22%), followed by college graduates (17%), persons with some college education (14%), and high school graduates (12%). Highest twitter usage was reported in urban areas (19%), followed by suburban areas (14%), and rural areas (8%). The demographics of the twitter population that would tweet about health related concerns, is unknown. Characteristics of twitter usage in relation to age, race, education level, and locality can affect the generalizability of findings.

V. CONCLUSION

Probability of “norovirus” hashtags occurring above the daily mean on a day with “fever” hashtags above daily mean were statistically significant. “Norovirus” hashtag had the highest correlation with “fever” hashtag, followed by “outbreak”, “throwing up”, and “sick” hashtags. “Fever” and “sick” keywords had a statistically significant difference in relation to utilization of “norovirus” hashtag. A non-linear regression equation, using a polynomial of degree six, can be formed for short term extrapolation of norovirus incidence. Twitter-based epidemic surveillance is an irreplaceable resource for early warning on emerging outbreaks because of their sensitivity, availability, convenience, and transparency.

REFERENCES

[1] A. J. Hall, M. Rosenthal, N. Gregoricus, S. A. Greene, J. Ferguson, O. L. Henao, *et al.*, "Incidence of Acute Gastroenteritis and Role of Norovirus, Georgia, USA, 2004-2005," *Emerging Infectious Diseases,* vol. 17, pp. 1381-1388, 2011.

[2] I. Barrabeig, A. Rovira, J. Buesa, R. Bartolomé, R. Pintó, H. Prellezo, *et al.*, "Foodborne norovirus outbreak: the role of an asymptomatic food handler," *BMC Infectious Diseases,* vol. 10, pp. 269-275, 2010.

[3] A. S. Chapman, C. T. Witkop, J. D. Escobar, C. A. Schlorman, L. S. DeMarcus, L. M. Marmer, *et al.*, "Norovirus outbreak associated with person-to-person transmission, U.S. Air Force Academy, July 2011," *Msmr,* vol. 18, pp. 2-5, 2011 2011.

[4] A. Yilmaz, K. Bostan, E. D. A. Altan, K. Muratoglu, N. Turan, D. Tan, *et al.*, "Investigations on the Frequency of Norovirus Contamination of Ready-to-Eat Food Items in Istanbul, Turkey, by Using Real-Time Reverse Transcription PCR," *Journal of Food Protection,* vol. 74, pp. 840-843, 2011.

[5] O. Zacheus and I. T. Miettinen, "Increased information on waterborne outbreaks through efficient notification system enforces actions towards safe drinking water," *Journal of Water and Health,* vol. 9, pp. 763-772, Dec 2011.

[6] J. C. M. Heijne, M. Rondy, L. Verhoef, J. Wallinga, M. Kretzschmar, N. Low, *et al.*, "Quantifying Transmission of Norovirus During an Outbreak," *Epidemiology,* vol. 23, pp. 277-284, Mar 2012.

[7] M. Summa, C.-H. von Bonsdorff, and L. Maunula, "Pet dogs—A transmission route for human noroviruses?," *Journal of Clinical Virology,* vol. 53, pp. 244-247, 2012.

[8] B. A. Lopman, A. J. Hall, A. T. Curns, and U. D. Parashar, "Increasing Rates of Gastroenteritis Hospital Discharges in US Adults and the Contribution of Norovirus, 1996-2007," *Clinical Infectious Diseases,* vol. 52, pp. 466-474, 2011.

[9] R. Fretz, D. Schmid, S. Jelovcan, R. Tschertou, E. Krassnitzer, M. Schirmer, *et al.*, "An outbreak of norovirus gastroenteritis in an Austrian hospital, winter 2006-2007," *Wiener Klinische Wochenschrift,* vol. 121, pp. 137-143, Feb 2009.

[10] L. Hualiang, N. Sammy, C. Shelley, C. Wai Man, K. C. K. Lee, S. C. Ho, *et al.*, "Institutional risk factors for norovirus outbreaks in Hong Kong elderly homes: a retrospective cohort study," *Bmc Public Health,* vol. 11, pp. 297-303, 2011.

[11] M. Wadl, K. Scherer, S. Nielsen, S. Diedrich, L. Ellerbroek, C. Frank, *et al.*, "Food-borne norovirus-outbreak at a military base, Germany, 2009," *BMC Infectious Diseases,* vol. 10, pp. 1-10, 2010.

[12] A. Doménech-Sánchez, C. Juan, J. L. Pérez, and C. I. Berrocal, "Unmanageable norovirus outbreak in a single resort located in the Dominican Republic," *Clinical Microbiology & Infection,* vol. 17, pp. 952-954, 2011.

[13] Centers for Disease Control and Prevention, "Multisite Outbreak of Norovirus Associated with a Franchise Restaurant -- Kent County, Michigan, May 2005," *MMWR: Morbidity & Mortality Weekly Report,* vol. 55, pp. 395-397, 2006.

[14] Reuters, "Hundreds on QE 2 Sick with Suspected Stomach Flu," in *Reuters*, ed, 2007.

[15] U. Parashar, E. S. Quiroz, A. W. Mounts, S. S. Monroe, R. L. Fankhauser, T. Ando, *et al.*, ""Norwalk-like viruses". Public health consequences and outbreak management," *MMWR. Recommendations and reports : Morbidity and mortality weekly report. Recommendations and reports / Centers for Disease Control,* vol. 50, pp. 1-17, 2001 Jun 2001.

[16] A. L. Greer, S. J. Drews, and D. N. Fisman, "Why "Winter" Vomiting Disease? Seasonality, Hydrology, and Norovirus Epidemiology in Toronto, Canada," *Ecohealth,* vol. 6, pp. 192-199, Jun 2009.

[17] E. Peter and M. Blake, "26,500 school cafeterias lack required inspections," ed.

[18] N. C. f. I. a. R. D. Division of Viral Diseases, Centers for Disease Control (CDC). (2012, March 18, 2012). *Norovirus* Available: http://www.cdc.gov/ncidod/dvrd/revb/gastro/norovirus.htm

[19] C. Chew and G. Eysenbach, "Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak," *Plos One,* vol. 5, Nov 2010.

[20] J. H. Jones and M. Salathe, "Early Assessment of Anxiety and Behavioral Response to Novel Swine-Origin Influenza A(H1N1)," *Plos One,* vol. 4, Dec 2009.

[21] G. Eysenbach, "Infodemiology: tracking flu-related searches on the web for syndromic surveillance," *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium,* pp. 244-8, 2006 2006.

[22] P. Kostkova, E. de Quincey, and G. Jawaheer, "The potential of social networks for early warning nad outbreak detection systems: the swine flu Twitter study," *International Journal of Infectious Diseases,* vol. 14, pp. E384-E385, Mar 2010.