# Tweeting Fever: Can Twitter Be Used to Monitor the Incidence of Dengue-Like Illness in the Philippines?

*Jacqueline S. Coberly, Clayton R. Fink, Yevgeniy Elbert, In-Kyu Yoon, John Mark Velasco, Agnes D. Tomayao, Vito Roque Jr., Enrique Tayag, Durinda R. Macasocol, and Sheri H. Lewis*

*The purpose of the pilot study described in this article was to investigate whether Twitter could be a viable data source for monitoring dengue-like illness in the Philippines. The results suggest that a relatively small but sufficient number of tweets mentioning dengue-like illness in a person can be isolated from data collected from the Twitter public application programming interface. More importantly, the temporal distribution of these dengue-like tweets was similar enough to the distribution of counts of new cases of dengue-like illness in the same region to suggest that the tweets could provide a valid data source for monitoring the temporal trend of dengue-like illness. Although it is not within the scope of the project described in this article, it would be relatively easy to operationalize the use of tweets collected from the Twitter public application programming interface as a timely, valid data source for an electronic disease surveillance system.*

## INTRODUCTION

Dengue, also known as "breakbone fever" for the severe myalgia and joint pain experienced by patients, is a major cause of morbidity and mortality around the world. It is caused by a *Flavivirus* that is transmitted to humans when they are bitten by infected mosquitoes.[1] There are four distinct serotypes of the dengue virus (DEN 1–4), all of which cause disease in humans that ranges from asymptomatic infection to severe, fatal hemorrhagic illness.[2] Recovery from infection provides serotype-specific immunity but does not protect from infection with other serotypes of the dengue virus. Rather, repeat infection with a different serotype may be associated with increased risk of severe hemorrhagic dengue.[3] The incidence of dengue has increased 30-fold since the first severe outbreaks of hemorrhagic dengue were recognized in the Philippines and Thailand in the 1950s.[4–6] This rate of increase classifies dengue as an emerging infection with the potential to cause a global epidemic or pandemic.[7] Indeed, dengue is now endemic in more than 100 countries around the world, putting nearly 50% of the global population at risk of infection.[8] Reports of new cases to the World Health Organization

suggest that 50–100 million people around the world contract dengue each year.[3] Many cases are subclinical, however, so a better estimate may be closer to 390 million infections per year, of which 96 million present clinically.[9] The recent and increasing occurrence of clusters of dengue in the southern United States and across Europe also suggests that the geographic distribution of the virus is spreading as global warming increases temperatures, creating potential mosquito habitats in formerly temperate zones.[10–13]

With no effective drug therapy or vaccine, control of the mosquito vector and surveillance for clinical infections are the primary public health tools available to fight dengue.[14] Early identification and location of outbreaks can help target intervention campaigns to reduce existing mosquito populations and breeding areas in high-risk locations. The goal of such campaigns is to minimize the spread and impact of an outbreak, but to be effective, intervention needs to start as soon as possible.

Public health authorities in many resource-rich countries use electronic disease surveillance systems to improve the timeliness of disease detection.[15–17] Electronic systems allow authorities to monitor the spread of disease in a population in near real-time fashion. These systems use computerized health data from multiple sources to generate displays of the frequency of new cases of disease temporally and geographically. They can improve early identification of potential outbreaks but only if the computerized data are available quickly—ideally, the day they are collected.

Electronic disease surveillance can be especially valuable in resource-limited areas where new infectious agents frequently arise, and electronic systems targeted for these environments are available.[18] Unfortunately, resource-limited countries, which include most dengue-endemic countries in the world, often lack the infrastructure and resources needed to rapidly digitize the health data needed for electronic disease surveillance. Medical data are often not computerized in these areas or are not computerized quickly enough to be useful in near-real-time surveillance. Although many resource-limited countries are moving toward electronic disease surveillance, implementation of data collection and data transmission protocols will take time and funding and is easily a decade or more from completion. In the meantime, other data sources are being sought that could be used now by an electronic disease surveillance system to monitor disease trends.

A number of electronic systems, such as BioCaster, HealthMap, Global Public Health Intelligence Network, and EpiSPIDER, mine publicly available electronic news media for reports of specific diseases.[19] Some of these systems have been in use for more than a decade and have provided useful information on disease trends. Unfortunately, news reports tend to lag behind an outbreak, so mining news reports may not provide information quickly enough for public health professionals to intervene and slow the spread of an outbreak. Social media, such as the microblogging platform Twitter, provides digitized data continuously 24 hours per day. Posts on Twitter, or tweets, are limited to 140 characters or less, and users tweet to update friends on their activities and thoughts. The content of tweets, therefore, varies wildly—from social commentary to what the user is having for dinner. Most tweets are publicly available through the Twitter application programming interface (API). A pseudo-random sample of tweets meeting user-specified criteria can be obtained relatively easily and free of cost from the Twitter API. Twitter is also heavily used in many resource-limited areas where other sources for electronic disease surveillance are limited. The Philippines, for example, is among the top 20 producers of tweets in the world.[20]

Multiple investigators have mined tweets for information about the behaviors, moods, and habits of Twitter users, and some have also looked for information to inform disease surveillance.[21–27] Investigators used Twitter to monitor influenza activity in the United States during the H1N1 pandemic in 2009–2010 and noted good correlation with the number of new influenza cases as collected by public health authorities.[25] Similarly, Collier et al. found a moderately strong association between World Health Organization/National Respiratory and Enteric Virus Surveillance System laboratory incidence data for influenza and the incidence of tweets mentioning influenza during the 2009–2010 influenza season in the United States.[26] Outside of the United States, Chunara et al. compared the volume of cholera reports for Haiti collected from HealthMap (http://www.healthmap.org) and Twitter posts with the number of new cholera cases collected via standard surveillance methods by the Haitian Ministry of Public Health. They found a statistically significant positive correlation between the combined HealthMap/Twitter data and the incidence of cholera as collected by the Haitian Ministry of Public Health data (Pearson correlation coefficients ranging from 0.76 to 0.86).[28] Another study by Chan et al. found significant positive correlations (Pearson correlation coefficients from 0.82 to 0.99) between the number of tweets mentioning "dengue" or similar phrases and dengue incidence as measured by public health authorities in Bolivia, Brazil, India, Indonesia, and Singapore.[29]

If a subset of tweets that mimics the true incidence (i.e., the count of new cases) of a disease in a population could be reliably identified, it would be relatively simple to set up a continuous feed of tweets from the Twitter API, process the raw tweets to extract the appropriate tweet subset, and feed those tweets directly into an electronic disease surveillance application. This would provide an inexpensive, yet timely, surrogate disease surveillance data source.

# METHODS

## Study Design

The study described in this article has two objectives: to determine whether tweets mentioning dengue-like illness in an individual can be identified in the Twitter sample collected; and if so, to determine whether the temporal distribution of these "dengue-like" tweets is similar enough to the temporal distribution of new counts of dengue-like illness, as collected by Philippines public health authorities, to be used as a data source to monitor dengue-like illness in the Philippines.

Under optimal conditions, a diagnosis of dengue fever is confirmed by laboratory tests that identify the presence of the dengue virus or antibodies to the virus in the blood of a patient. These blood tests are not always available, however, and in their absence dengue is diagnosed clinically, based on the presentation of a specific set of symptoms in a patient. The clinical diagnosis of dengue used in the Philippines in 2011 was a patient presenting with fever and one or more of the following symptoms: headache, eye pain, muscle or joint pain, rash, nausea, or vomiting. The cases discussed in this article include both those confirmed to be dengue by a laboratory test and those diagnosed clinically by a physician; therefore the term *dengue-like illness* is used instead of *dengue*.

## Tweet Collection

Tweets were collected using Version 1.0 of the free Twitter public API, which allows an individual to request a feed of public tweets matching specific search criteria. Each request, or query, returns a 1% pseudo-random sample of all tweets meeting those criteria, although the precise tweet selection process used by the API has not been disclosed by Twitter. Two separate search criteria were used to collect tweets for this study. The first API query asked for tweets from two areas of the Philippines for specified time periods: 18 June 2011 through 9 September 2011 for Cebu City, Philippines, and 24 July 2011 through 16 September 2011 for the National Capital Region (NCR), which includes Manila and surrounding suburbs. The second API query requested all tweets from the Twitter users whose tweets were returned by the first geographic query. Tweets from both API queries were combined for this analysis.

For tweet collection, Cebu City and the NCR were defined geographically by the latitude and longitude of a point at the center of the region and a radius in miles extending out from the central point (Table 1). The location of a tweet was recorded as the latitude and longitude of the tweeting device if geotagging was enabled on the device. For geotagged tweets, the latitude and longitude were extracted from the tweet metadata, and the closest populated place, as based on a lookup against an online gazetteer (GeoNames, www.geonames.org), was taken as the user's location. If geotagging was not enabled, the user's location was inferred by matching the location given in his or her Twitter profile against the gazetteer. Only tweets that mapped to a location within the specified geographic coordinates in Table 1 were retained.

## Identification of Tweet Subsets

The Twitter convention *@username* was used to identify and remove usernames to anonymize the tweets. Duplicate tweets and retweets (tweets posted by one user and then forwarded by another user)[30] were removed from the data set before analysis. During preliminary examination of tweets, several words commonly included in tweets not containing mention of dengue-like illness were identified and the corresponding tweets were removed before analysis.

### *Simple Keyword Searches for Dengue-Like Tweet Subsets*

Although only a fraction of all the tweets mentioned the keyword *fever*, it is the only required symptom in the clinical case definition used in the Philippines, and public health authorities in Cebu City, Philippines, monitor new reports of undifferentiated fever as a surrogate measure of dengue-like illness in seasonal surveillance activities. For those reasons, this term was chosen as the focus of the simple keyword analysis. The keywords *fever* and *feverish* were examined. Dengue-like (DL) tweet subsets were created by searching tweets for those keywords in English and/or Tagalog, the native language of the Philippines. For the Fever subset, tweets containing

**Table 1. Descriptive statistics of tweet data**

| Variable | Cebu City | NCR | Total |
|---|---|---|---|
| Central (latitude, longitude) | (10.31667, 123.95) | (14.63333, 121.03333) | — |
| Radius (miles) of tweet locations | 15 | 20 | — |
| Dates of tweet collection | 18 June–9 September 2011 | 24 July–16 September 2011 | 18 June–16 September 2011 |
| No. of people who tweeted | 31,015 | 137,281 | 168,296 |
| Total no. of tweets collected | 3,769,746 | 11,981,026 | 15,750,772 |

*fev* or *lagnat* were selected, and tweets containing the words *feverish* or *nilalagnat* or *may sinat* or *sinisinat* were used to identify the Feverish Tweet subset. Each subset is labeled with the keyword it represents: Fever DL Tweet and Feverish DL Tweet.

bodyache, rashes, spasms, malaise, abdominal, ~~fever~~, nausea, ~~lagnat~~, headache, vomiting, stomachache, throat, ~~scarlet~~, muscle, symptoms, sore, chills, slight, rash, joint, dengue, aches, bleeding, ache, pain, blood, meds, clinic, ~~38~~, stomach, body, painful

**Figure 1.** Expansion word candidates in descending order of PMI.

### Human-Tagged Tweet Subset

The Fever DL Tweets were reviewed manually to identify those actually mentioning a person who was sick with a fever, with or without other symptoms. Tweets that used the word *fever* in a nonclinical way were excluded. These tweets were tagged and form the human-tagged tweet subset (HT Tweet).

### Using Query Expansion to Retrieve DL Tweets

All tweets were indexed using the Lucene text indexing and search API (http://lucene.apache.org/). High-frequency words such as articles, pronouns, and prepositions, in both English and Tagalog, were removed. URLs were converted to the tag *_url_*. The high-precision query used was: [(fever lagnat) AND (headache rash pain bleed* blood) NOT cold* NOT cough* NOT nose NOT _url_ NOT bieb*]. In the query, * is used to indicate that any word starting with the preceding string should be matched. This query asks for tweets that contain the word *fever* and one or more of the words *headache*, *rash*, *pain*, *bleed**, or *blood*. It also adds the restriction that the tweet must not contain any of the words *cold**, *cough**, or *nose*; must not contain a URL; and must not contain references to Justin Bieber (i.e., _NOT bieb*). These restrictions were used to eliminate respiratory complaints and tweets containing links to health-related sites. Next, words that were most closely associated with the results of this query as compared to all tweets in the index were identified by calculating the normalized pointwise mutual information (PMI) for each word returned in the query results but excluding the original query terms where $N$ equals all tweets:

$$pmi(x;y) = \frac{p(x,y)}{p(x)p(y)}$$

$$nmpi(x;y) = \frac{pmi(x;y)}{-\log[p(x,y)]}$$

$$p(x,y) = \frac{|\,query\ results\ containing\ word\ y\,|}{N}$$

$$p(x) = \frac{|\,query\ results\,|}{N}$$

$$p(y) = \frac{|\,all\ tweets\ containing\ word\ y\,|}{N}$$

PMI is an information theoretic measure of association between two random variables. The normalized form of PMI maps the values to the range (–1 to 1), where –1 means no association, 0 reflects total independence, and 1 represents complete association. The random variables in this case are the number of tweets matching the high-precision query, $x$, and the number of tweets, $y$, in the complete index that contain a word that was contained in three or more of the returned tweets. These words form the set of words that co-occur with the query terms. Words that are more likely than not to co-occur with query terms will have a high PMI. The top 32 words, as scored by their calculated PMI, are shown in Fig. 1. Words with a strike-through were not used as expansion terms. The expansion terms were then added as a disjunction "AND-ed" to a query for *fever or lagnat* and run against the index again to retrieve an expanded set of tweets.

### Dengue Incidence Data

Two sources of daily counts of new dengue-like cases of illness were used in this study: counts of dengue and dengue-like cases reported to the Philippines Integrated Disease Surveillance and Response System (PIDSR), and daily counts of the number of people presenting with fever at government-funded clinics in Cebu City, Philippines. PIDSR is an integrated disease surveillance system used throughout the Philippines to collect information about nationally reportable diseases.[31] Incidence data for selected diseases are collected and summarized at the local level and sent forward through municipal and provincial public health authorities to the National Epidemiological Center (NEC) where surveillance data are compiled for the whole country.[31] Use of anonymized PIDSR data for individual patients from the NCR and Cebu City in this study was approved by the NEC. As in other notifiable disease systems around the world, illnesses reported in PIDSR are thoroughly investigated so receipt of the information at NEC is often delayed. The date of disease onset and the date the case is reported are both included in each report, and the onset date was used to plot cases temporally for this analysis. PIDSR data were available for all of 2011, but only data from 8 June 2011 through 26 September 2011 were used. This period corresponds to the dates when tweets were collected plus an additional 10 days at the beginning and end of the period, allowing examination of the effect of shifting tweets forward and backward in time on the

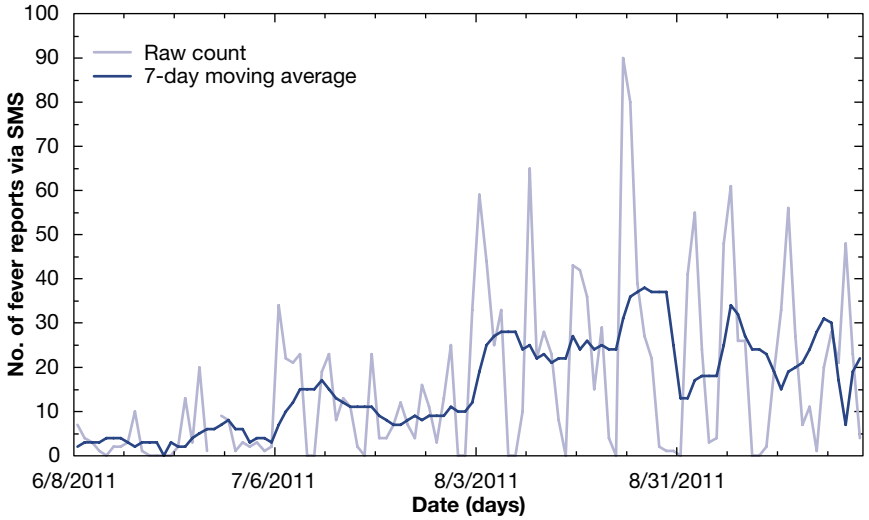


**Figure 2.** Raw and smoothed counts of SMS fever reports for Cebu City by date.

correlation with the incidence data. To address the disproportionate sampling of cases from Cebu City and the NCR, Pearson correlation coefficients were computed to compare the temporal distribution of cases between the two locations at different points in time.

The second source of incidence data is unique to Cebu City, Philippines. The Cebu City Health Department (CHD), the public health authority for the city, has traditionally used the number of new cases of undifferentiated fever reported by government health clinics each day as a simple way to track dengue-like illness during the peak dengue season in May through December (personal communication, D. Macasoco, CHD).

In 2009, the CHD replaced its paper-based fever system with an electronic system that collects data via short message service (SMS) cellular phone messages.[32] Each day, personnel at government clinics throughout the city send a single SMS to a dedicated phone line at the CHD for each person presenting at the clinic with fever. The SMS messages are received by a mobile phone attached to the dedicated line and are automatically transferred to a desktop computer. A custom application on the computer receives and parses the SMS for validity and stores valid messages in a database. Fever time series compiled from these data are reviewed to monitor fever incidence in Cebu City. The date of onset of fever is not included in the fever SMS data, so the date of the clinic visit is used to plot these cases temporally. The valid fever SMS messages (Fever SMS) from this system for June to November 2011 were provided by the CHD for this analysis, and as with the PIDSR data, only the data from 8 June 2011 through 26 September 2011 were used. Because the government clinics in Cebu City do not generally see patients on weekends, the Fever SMS data show a strong day-of-week effect that is not seen in the PIDSR or Twitter data because they are reported or collected daily. To facilitate comparison of the Fever SMS to the PIDSR and Twitter data, a 7-day moving average of the Fever SMS counts was computed, and the resulting daily averages, rounded to the nearest whole number, were used when comparing the Fever SMS data to other data for temporal correlation (Fig. 2).

### *Comparison of HT and Other Tweet Subsets with Dengue Incidence Data*

The temporal distribution of the HT Tweet, Fever DL Tweet, Feverish DL Tweet, and PMI Tweet subsets were compared to the temporal distribution of the Fever SMS average counts and the PIDSR counts of new cases of dengue-like illnesses. The Fever, Feverish, and PMI Tweet subsets were also compared temporally to the HT Tweet subset. Pearson correlation coefficients were computed as a measure of agreement for the different comparisons.

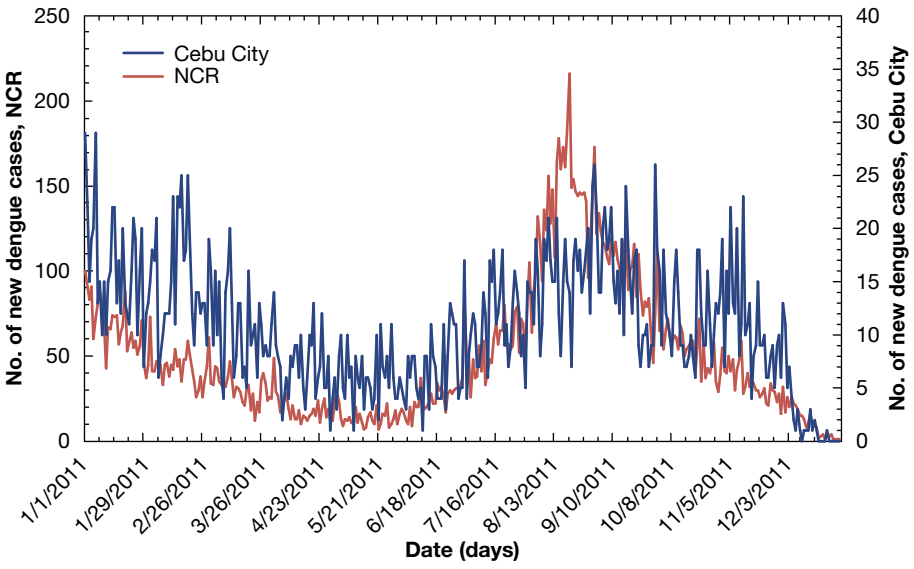


**Figure 3.** Number of cases of dengue reported in PIDSR in 2011 by location and time.

**Table 2. Pearson correlation coefficients by location and time for Fever SMS, PIDSR, and all Fever Tweets**

| Subset | 18 June–23 July 2011 | 24 July–16 September 2011 | 18 June–16 September 2011 |
|---|---|---|---|
| Cebu City PIDSR vs. NCR PIDSR | 0.474 | 0.303 | 0.598 |
| SMS vs. Cebu City PIDSR | 0.439 | 0.198 | 0.533 |
| SMS vs. Cebu City and NCR PIDSR | 0.464 | 0.504 | 0.775 |
| Cebu City Fever DL Tweets vs. SMS | 0.625 | n/a | n/a |
| NCR Fever DL Tweets vs. SMS | n/a | 0.164 | n/a |
| Cebu City Fever DL Tweets vs. Cebu PIDSR | 0.393 | n/a | n/a |
| NCR Fever DL Tweets vs. NCR PIDSR | n/a | 0.107 | n/a |

To examine whether the tweet subsets might provide dengue-like illness trend information more or less quickly than Fever SMS or PIDSR counts, the tweet subsets were shifted forward and backward in time, day by day, for up to 10 days each way, and the correlation of the shifted data to the unshifted Fever SMS and PIDSR counts was recomputed for each of the daily shifts.

## RESULTS

### Description of PIDSR and SMS Data

The incidence of PIDSR data from Cebu City and the NCR was compared to see whether the temporal patterns of reported cases of dengue-like illness were similar in the two cities in 2011 (Fig. 3). The correlation between the cities was positive and statistically significant, albeit moderate (Pearson correlation coefficient, 0.598, $p < 0.001$). Comparisons were also made for the time period when tweets were collected solely in Cebu City (18 June 2014 through 23 July 2014) and for the period after the start of collection of tweets from the NCR, 24 July 2014 through 16 September 2014). Correlation in both periods was somewhat lower than in the combined period but still positive (Table 2). Given the general similarity in distribution of dengue-like case reports from the two cities, the data were combined during comparisons to the tweet data sets.

The correlation of the SMS (collected only in Cebu City) and Cebu City PIDSR data was moderate but positive and statistically significant (0.533, $p < 0.001$), validating the use of the Fever SMS data as a surrogate for dengue-like illness in public health disease surveillance activities (Table 2). The correlation of the SMS data to the combined Cebu City plus NCR PIDSR data was also positive (0.775, $p < 0.001$) and increased when the SMS were shifted to the left by six days (0.826, $p < 0.001$) (Fig. 4). This lag is expected because the PIDSR data are measured from date of onset and the SMS data from clinic visit, which logically follows the date of onset. Because the PIDSR data have an average 14.8-day lag from onset to data entry, however, the SMS data likely provide timelier trend data.

### Description of Tweets and Creation of Tweet Subsets

A total of 15,750,771 tweets were collected prospectively from 18 June 2011 through 16 September 2011

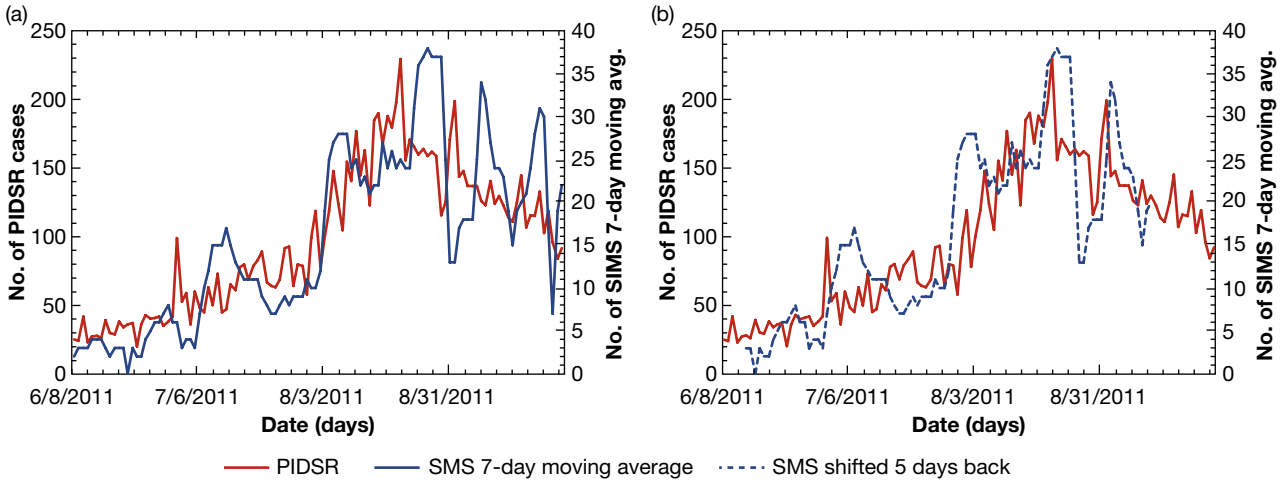


**Figure 4.** Temporal distribution of PIDSR reports vs. SMS reports and SMS reports shifted –5 days.

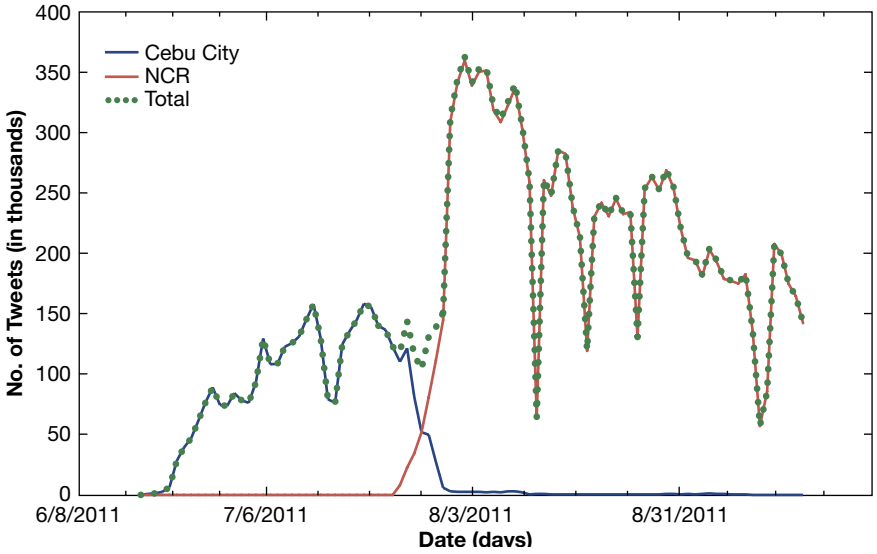


**Figure 5.** All tweets collected by location and date.

(Table 1). Initially tweets were collected only from Cebu City, but returns were relatively low. To augment the small number of tweets being collected from Cebu City, the Twitter API request was changed in late July 2011 to add tweets from the NCR. This process increased the total number of tweets collected. It also decreased the number of tweets collected from Cebu City, however, because the 1% sample of tweets returned by the Twitter API was then split between Cebu City and the NCR. Because the NCR has a much larger population, the Cebu City tweets were grossly reduced after 23 July 2011 when tweet collection began in the NCR (Fig. 5). By 1 August 2011, Cebu City contributed 1% or less of the tweets collected each day. To compensate for the decline in the Cebu City Twitter feed, the tweets from the two locations were combined during most analyses.

The content of the tweets varied wildly. Nearly a quarter (3,849,264) of the tweets were exact duplicates or retweets (Fig. 6). Review of tweets containing the term *fever* showed that *fever* had multiple meanings in the tweets. Some tweets did mention fever as a symptom of an illness in a person, but it was used most often to describe obsessive activity or strong emotions. For example, 127,958 (0.8%) of all tweets proclaimed [Justin] Bieber fever, [Harry] Potter fever, [David] Azkal fever, or a fever for some other person or place. In addition, tweets containing the term *ha* or *ha ha* (4,399,242 tweets, or 27.9%) generally meant that the word *fever* was being used in a joking fashion rather than as a description of illness. Removal of the *I have a fever for . . .* and the joking tweets left a total of 7,424,308 tweets. Of those, 6,235 contained the word *fever* (Fig. 6), and these tweets make up the Fever DL Tweet subset.

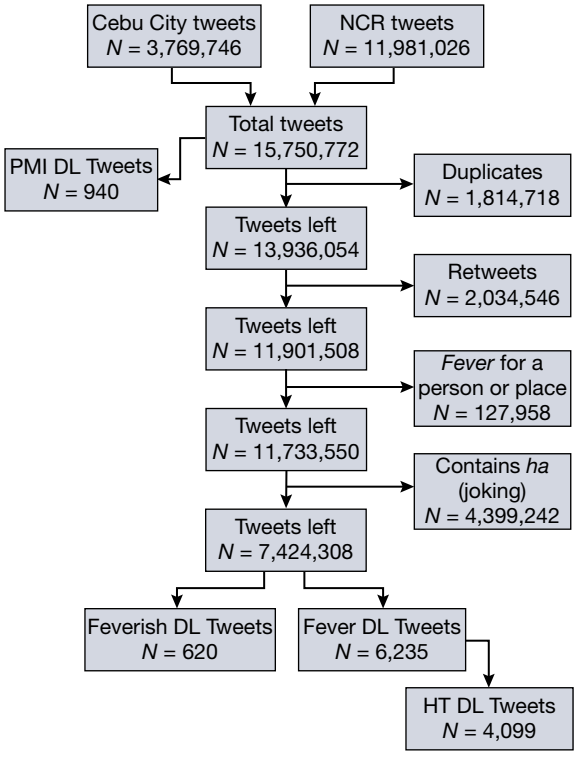


**Figure 6.** Creation of DL tweet subsets.

The Fever DL Tweet subset was reviewed manually to identify tweets that, in fact, used the term *fever* to describe a person with a dengue-like illness. A total of 4,099 tweets met that definition and are included in the HT Tweet subset. A similar query of the refined tweet set ($N$ = 7,424,308) for tweets containing the English and Tagalog words for *feverish* produced 620 tweets that make up the Feverish DL Tweet subset. The more complex keyword query that used the PMI calculations was applied to the initial data set to create the PMI Tweet subset containing 940 tweets.

## Correlation of DL Tweet Subsets with Fever SMS and PIDSR Counts

Because the HT Tweets are a subset of the Fever DL Tweets, a positive correlation was expected and observed between the HT Tweets and the Fever and Feverish Tweet subsets (Table 3). Shifting the Fever and Feverish Tweet subsets in time did not increase their correlation with the HT Tweets.

**Table 3. Pearson correlation coefficients for pairs of DL subsets vs. incidence counts**

| Subset | Fever DL Tweets | Feverish DL Tweets | PMI Tweets | HT Tweets | Fever SMS | PIDSR |
|---|---|---|---|---|---|---|
| Fever DL Tweet | 1.0 | 0.849 | 0.761 | 0.920 | 0.611 | 0.601 |
| Feverish DL Tweet | | 1.0 | 0.701 | 0.849 | 0.541 | 0.552 |
| PMI Tweet | | | 1.0 | 0.858 | 0.721 | 0.746 |
| HT Tweet | | | | 1.0 | 0.658 | 0.712 |
| Fever SMS | | | | | 1.0 | 0.786 |
| PIDSR | | | | | | 1.0 |

**Table 4. Maximum Pearson correlation coefficients for time-shifted DL subsets vs. incidence counts**

| Subset | SMS [Time Shift, Days] | PIDSR [Time Shift, Days] |
|---|---|---|
| Fever Tweet | 0.679 [+3] | 0.764 [+9] |
| Feverish Tweet | 0.613 [+4] | 0.735 [+7] |
| PMI Tweet | 0.721 [+0] | 0.752 [+5] |
| HT Tweet | 0.745 [+6] | 0.819 [+9] |

The HT Tweets were also positively correlated with both the Fever SMS and the combined PIDSR incidence counts (Pearson correlation coefficients, 0.658 and 0.712, respectively, $p < 0.001$) (Table 3). Correlation with both sets of incidence counts increases when the HT Tweets are shifted forward in time, increasing the correlation to 0.745 (+6 days) for Fever SMS and 0.819 (+9 days) for the PIDSR data (Table 4 and Fig. 7). This suggests the HT Tweets could provide information on changes in the trend of dengue-like illness nearly a week earlier than the two traditional sources of dengue incidence data. If the 14.8-day average lag-time between disease onset and data entry (i.e., when the PIDSR data are ready for analysis) is included for PIDSR, the HT Tweets could lead the PIDSR data by as much as 3 weeks.

The Fever and Feverish DL Tweet subsets also showed statistically significant positive correlations with the Fever SMS and combined PIDSR incidence data counts

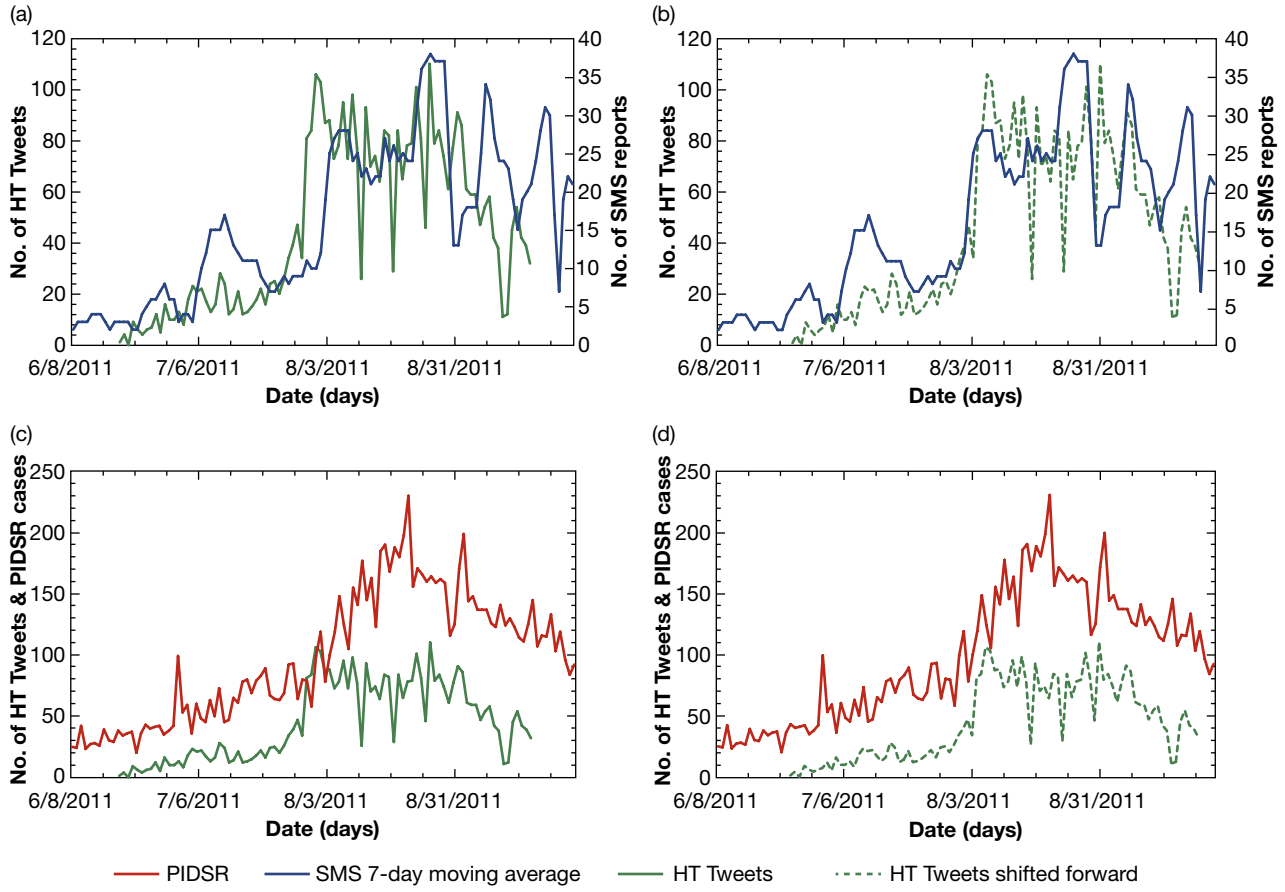


**Figure 7.** Unshifted and shifted HT Tweets vs. Fever SMS 7-day moving average counts and PIDSR counts. (a) HT Tweets vs. SMS; (b) shifted HT Tweets vs. SMS; (c) HT Tweets vs. PIDSR; (d) shifted HT Tweets vs. PIDSR.

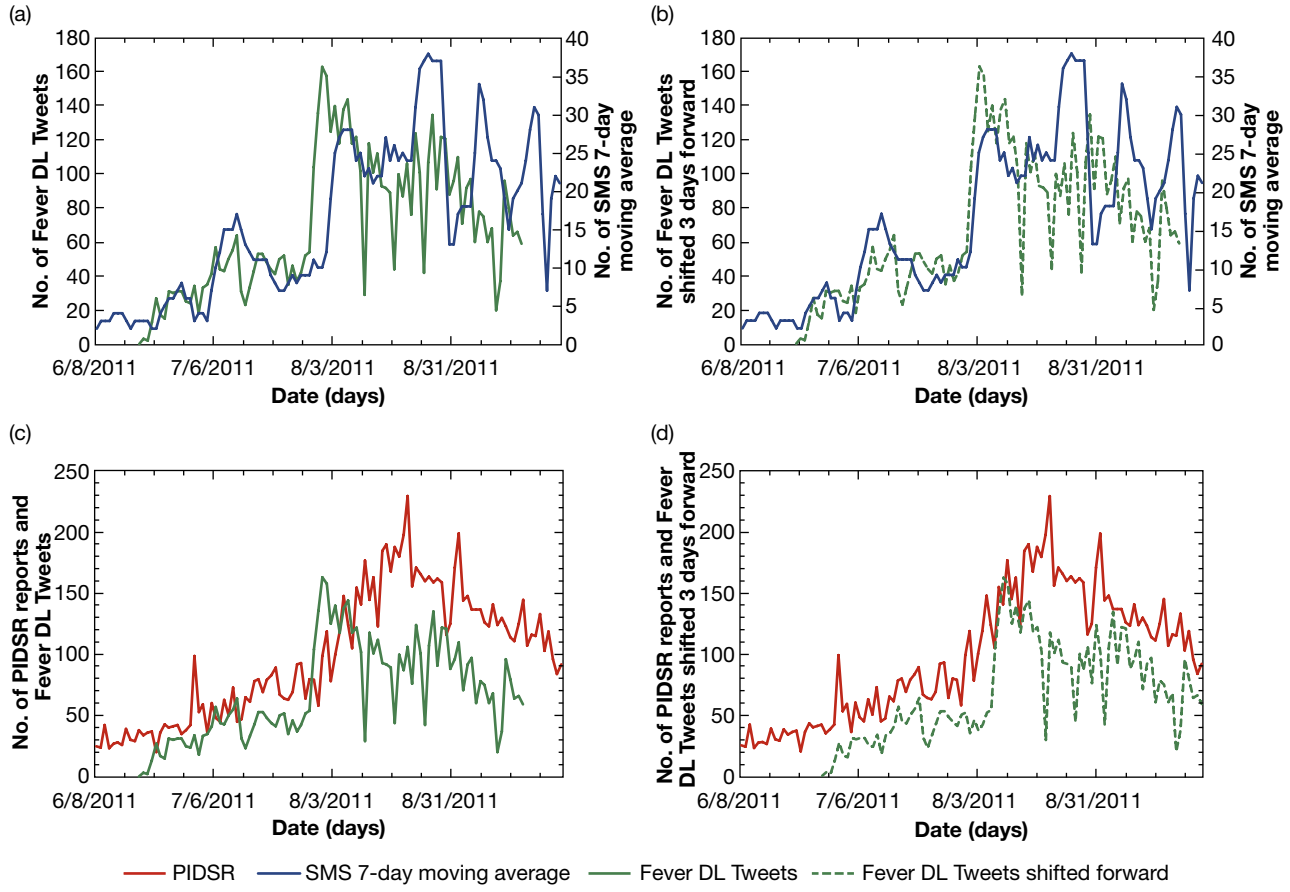


**Figure 8.** Unshifted and shifted Fever DL Tweets vs. Fever SMS 7-day moving average counts and PIDSR counts. (a) Fever DL Tweets vs. SMS; (b) shifted Fever DL Tweets vs. SMS; (c) unshifted Fever DL Tweets vs. PIDSR; (d) shifted Fever DL Tweets vs. PIDSR.

(Table 3). The correlations of the Fever and Feverish DL Tweets with the Fever SMS and PIDSR counts were weaker than those observed for the HT Tweet subset. As with the HT Tweets, the correlations of the Fever and Feverish DL Tweets with the Fever SMS and PIDSR counts were strengthened by moving the tweet subsets forward in time (Table 4), suggesting that the Fever and Feverish DL Tweet subsets could also provide earlier warning of changes in trends of dengue-like illness than the Fever SMS and the PIDSR incidence counts (Fig. 8). The correlation of the Fever DL Tweets was also examined by location. The correlations of the Fever DL Tweets with the SMS and PIDSR incidence counts remained positive for both Cebu City and the NCR locations, although the strength of the correlations decreased (Table 2).

The unshifted PMI Tweet subset also showed a positive correlation with the Fever SMS and PIDSR incidence data (0.7207 and 0.7464, respectively, $p < 0.0001$) (Table 3). The correlation of the PMI Tweets with PIDSR data increased when the tweets were shifted forward in time, but shifting the tweets in time did not improve correlation with the SMS data (Fig. 9). The correlation of the unshifted PMI Tweets with the SMS and PIDSR data is stronger than similar correlations observed for the Fever and Feverish DL Tweet subsets, but this advantage was reduced when the DL Tweet subsets were time-shifted (Table 4).

## DISCUSSION

This study identified several keyword-based methods used to isolate tweets from users in two locations in the Philippines who mention dengue-like illness in a person. The study showed that the temporal distribution of those tweet subsets is similar to the temporal distribution of counts of new dengue-like illness as recorded by Philippines public health authorities. Although the results are encouraging, this study was an exploratory pilot study with several limitations. The study addressed only a single disease in one country. To use Twitter as a source for electronic disease surveillance, the same preliminary work needs to be repeated for each illness monitored from Twitter data. Although using keyword permutations of the term *fever* was successful in the Philippines, the keywords will vary by location, if only because of language differences. The keyword distribution may also change over time, so ongoing evaluation

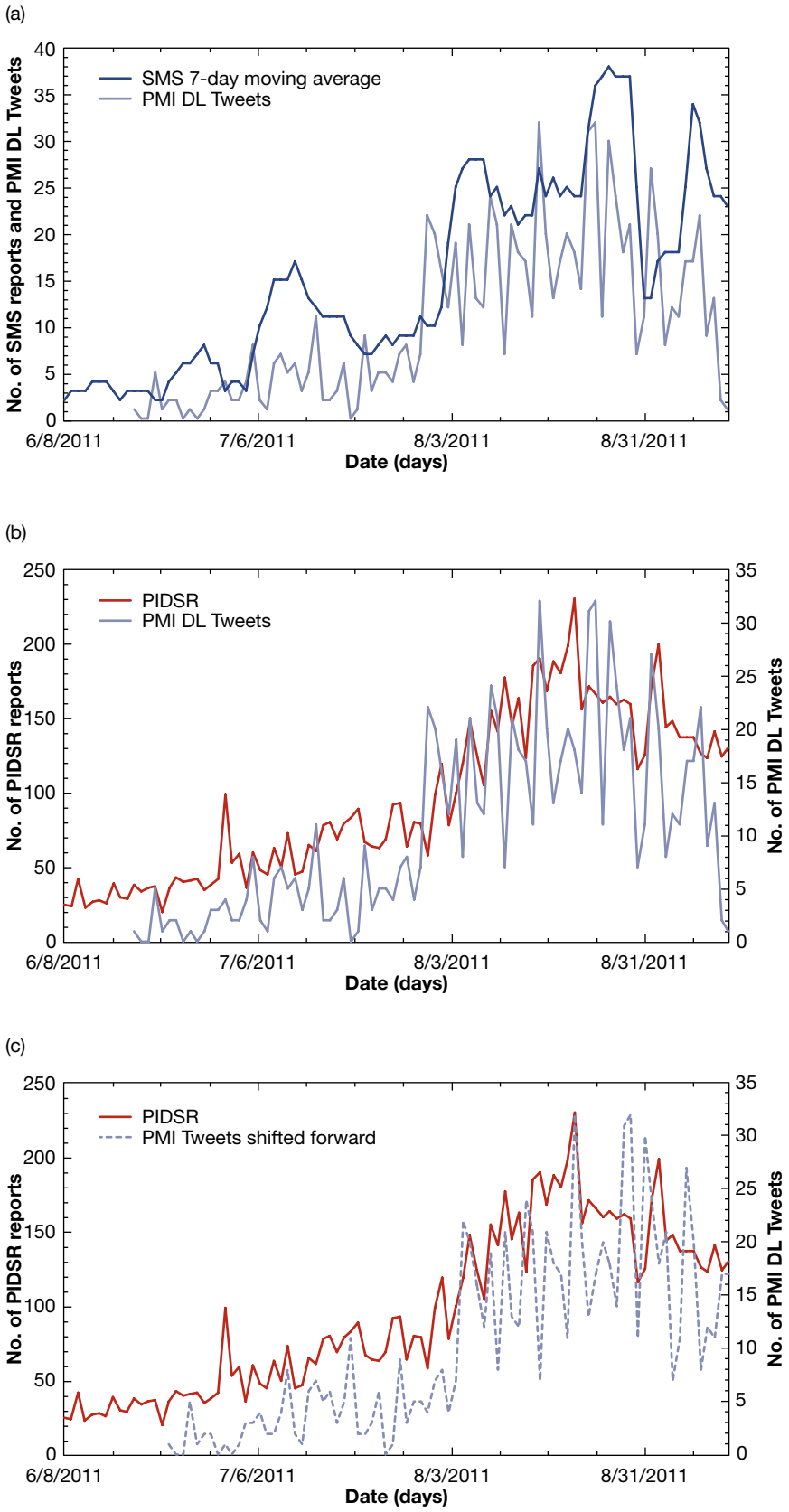


**Figure 9.** Unshifted and shifted PMI DL Tweets vs. Fever SMS 7-day moving average counts and PIDSR counts. (a) PMI DL Tweets vs. SMS; (b) PMI DL Tweets vs. PIDSR; (c) shifted PMI DL Tweets vs. PIDSR.

and update of the keyword set(s) would be needed if the tweets were used for disease surveillance long term.

There are also questions about the repeatability of data collected from the Twitter public API. The data from the API is, presumably, a pseudo-random 1% sample of tweets identified by the API queries used, but Twitter has not disclosed the exact methods used to create the 1% sample.[33] It is, therefore, possible that the outcome of sampling will vary by location, within a given location, over time, or by all these factors. This problem needs further evaluation before the API is used routinely in disease surveillance, because the cost of obtaining a larger ongoing Twitter feed is prohibitive in resource-limited areas.

The most serious limitation of this project was the decline in Cebu City tweets due to procedural changes in tweet collection. Separate analysis of the tweets by location show similar but weaker correlations to the Fever SMS and PIDSR data, suggesting that combination of tweets from Cebu City and the NCR did not bias the study results.

Adding a Twitter data feed to an electronic disease surveillance system would be relatively easy. Customized code would need to be written to capture the continuous feed of data from the free Twitter API for the appropriate geographic areas and then to write it to a database. Tweets stored in the database would need to be processed automatically to isolate those mentioning the specific illness being monitored. Hand-tagging the tweets containing information on fever would be too cumbersome for this process, but automated processes like those used to create the Fever DL and PMI DL Tweet subsets could be easily adapted for this purpose. Last, the tweets mentioning the specific illness would need to be visualized in an electronic disease surveillance system.

The primary advantages of using the Twitter data in disease surveillance are speed and cost; the Twitter public API is freely available and provides near-real-time computerized data. This type of surrogate data would augment, not replace, traditional public health disease surveillance. Its purpose is to help public health personnel identify and intervene in disease events rapidly, hopefully limiting the impact of the event. Traditional disease reporting is still needed to provide detailed, specific information on the incidence and movement of disease through a population, and the public health community must continue to move toward fully automated disease surveillance. Although further evaluation is clearly needed, this study suggests that the Twitter public API could provide a free source of disease incidence data for use in electronic disease surveillance.

**ACKNOWLEDGMENTS:** We thank Howard Burkom for his assistance in analysis, as well as the personnel at Philippines AFRIMS Virology Research Unit. The opinions or assertions in this article are the private views of the authors and do not necessarily reflect the official policy or position of the U.S. Department of the Army, the U.S. Department of Defense, or the U.S. government.

## REFERENCES

[1]Public Health Agency of Canada, "Dengue Fever Virus (DEN 1, DEN 2, DEN 3, DEN 4) – Material Safety Data Sheets (MSDS)," http://www.phac-aspc.gc.ca/lab-bio/res/psds-ftss/msds50e-eng.php (last modified 18 Feb 2011).

[2]New York State Department of Health, "Dengue Fever (Breakbone Fever, Dengue Hemorrhagic Fever)," http://www.health.ny.gov/diseases/communicable/dengue_fever/fact_sheet.htm (revised Oct 2011).

[3]Simmons, C. P., Farrar, J. J., van Vinh Chau, N., and Wills, B., "Dengue," *N. Engl. J. Med.* **366**(15), 1423–1432 (2012).

[4]Gubler, D. J., "Epidemic Dengue/Dengue Hemorrhagic Fever as a Public Health, Social and Economic Problem in the 21st Century," *Trends Microbiol.* **10**(2), 100–103 (2002).

[5]Guebler, D. J., Clark, G. G., "Dengue/Dengue Hemorrhagic Fever: The Emergence of a Global Health Problem," *Emerg. Infect. Dis.* **1**(2), 55–57 (1995).

[6]Kyle, J. L., Harris, E., "Global Spread and Persistence of Dengue," *Annu. Rev. Microbiol.* **62**, 71–92 (2008).

[7]World Health Organization, "Dengue Control," http://www.who.int/denguecontrol/en/index.html (accessed 21 Aug 2013).

[8]Centers for Disease Control and Prevention, "Dengue Epidemiology, http://www.cdc.gov/dengue/epidemiology/index.html (last updated 28 Oct 2013).

[9]Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., et al., "The Global Distribution and Burden of Dengue," *Nature* **496**(7446), 504–507 (2013).

[10]Añez, G., and Rios, M., "Dengue in the United States of America: A Worsening Scenario?" *Biomed. Res. Int.* **2013**, 678645 (2013).

[11]Hsieh, Y.-H., and Chen, C. W., "Turning Points, Reproduction Number, and Impact of Climatological Events for Multi-wave Dengue Outbreaks," *Trop. Med. Int. Health* **14**(6), 628–638 (2009).

[12]Griffiths, P., "Viruses in the Era of Global Warming," *Rev. Med. Virol.* **18**(2), 69–71 (2008).

[13]Khasnis, A. A., Nettleman, M. D., "Global Warming and Infectious Disease," *Arch. Med. Res.* **36**(6), 689–696 (2005).

[14]Rodriguez-Roche, R., and Gould, E. A., "Understanding the Dengue Viruses and Progress towards their Control," *Biomed. Res. Int.* **2013**, 690835 (2013).

[15]May, L., Chretien, J.-P., and Pavlin, J. A., "Beyond Traditional Surveillance: Applying Syndromic Surveillance to Developing Settings—Opportunities and Challenges," *BMC Public Health* **9**, 242 (2009).

[16]Patterson-Lomba, O., Van Noort, S., Cowling, B. J., Wallinga, J., Gomes, M. G., et al., "Utilizing Syndromic Surveillance Data for Estimating Levels of Influenza Circulation," *Am. J. Epidemiol.* **179**(11), 1394–1401 (2014).

[17]Samoff, E., Fangman, M., Hakenewerth, A., Ising, A., and Waller, A., "User of Syndromic Surveillance at Local Health Departments: Movement toward More Effective Systems," *J. Public Health Manag. Pract.* **20**(4), E25–E30 (2014).

[18]Borchert, J. N., Tappero, J. W., Downing, R., Shoemaker, T., Behumbiize, P., et al., "Rapidly Building Global Health Security Capacity—Uganda Demonstration Project, 2013," *MMWR Morb. Mortal. Wkly. Rep.* **63**(4), 73–76 (2014).

[19]Collier, N., Doan, S., Kawazoe, A., Goodwin, R. M., Conway, M., et al., "BioCaster: Detecting Public Health Rumors with a Web-based Text Mining System," *Bioinformatics* **24**(24), 2940–2941 (2008).

[20]Nico, "4 Ways How Twitter Can Keep Growing," *Peerreach blog*, blog.peerreach.com/2013/11/4-ways-how-twitter-can-keep-growing/ (7 Nov 2013).

[21]Lyon, A., Nunn, M., Grossel, G., and Burgman, M., "Comparison of Web-based Biosecurity Intelligence Systems: BioCaster, EpiSPIDER and HealthMap," *Transbound. Emerg. Dis.* **59**(3), 223–232 (2010).

[22]Keller, M., Blench, M., Tolentino, H., Freifeld, C. C., Mandl, K. D., et al., "Use of Unstructured Event-based Reports for Global Infectious Disease Surveillance," *Emerg. Infect. Dis.* **15**(5), 689–695 (2008).

[23]Freifeld, C. C., Mandl, K. D., Reis, B. Y., and Brownstein, J. S., "HealthMap: Global Infectious Disease Monitoring through Automated Classification and Visualization of Internet Media Reports," *J. Am. Med. Inform. Assoc.* **15**(2), 150–157 (2008).

[24]Dodds, P. S., Harris, K. D., Kloumann, I. M., Bliss, C. A., and Danforth, C. M., "Temporal Patterns of Happiness and Information in a Global Social Network: Hedonometrics and Twitter," *PLoS One* **6**(12), e26752 (2011).
[25]Signorini, A., Segre, A. M., and Polgreen, P. M., "The Use of Twitter to Track Levels of Disease Activity and Public Concern in the U.S. during the Influenza A H1N1 Pandemic," *PLoS One* **6**(5), e19467 (2011).
[26]Collier, N., Son, N. T., and Nguyen, N. M., "OMG U Got Flu? Analysis of Shared Health Messages for Bio-surveillance," *J. Biomed. Semant.* **2**(s5), s9 (2011).
[27]Cunningham, J. A., "Using Twitter to Measure Behavior Patterns," *Epidemiology* **23**(5), 764–765 (2012).
[28]Chunara, R., Andrews, J. R., and Brownstein, J. S., "Social and News Media Enable Estimation of Epidemiological Patterns Early in the 2010 Haitian Cholera Outbreak," *Am. J. Trop. Med. Hyg.* **86**(1), 39–45 (2012).
[29]Chan, E. H., Sahai, V., Conrad, C., and Brownstein, J. S., "Using Web Search Query Data to Monitor Dengue Epidemics: A New Model for Neglected Tropical Disease Surveillance," *PLoS Negl. Trop. Dis.* **5**(5), e1206 (2011).
[30]Tamura, Y., and Fududa, K., "Earthquake in Japan," *Lancet* **377**(9778), 1652 (2011).
[31]World Health Organization, "Asia Pacific Strategy for Emerging Diseases (APSED)," http://www.wpro.who.int/philippines/areas/surveillance_response/apsed/story_on_surveillance_riskassessment_response/en/index.html (accessed 28 Aug 2013).
[32]Coberly, J., Wojcik, R., Tomayao, A. D., Tac-an, I. A., Velasco, J. M. S., and Lewis, S., "Dengue SMS Surveillance Project in the Philippines," *Am. J. Trop. Med. Hyg.* **81**(5 Suppl 1), 289 (2009).
[33]Morstatter, F., Pfreffer, J., Liu, H., and Carley, K. M., "Is the Sample Good Enough? Comparing Data from Twitter's Streaming API with Twitter's Firehose," in *Proc. 7th International AAAI Conf. on Weblogs and Social Media*, Boston, MA, pp. 1–10 (2013).

# The Authors

**Jacqueline S. Coberly** is an APL infectious disease epidemiologist who applies her academic and field experience to the development, implementation, and evaluation of electronic disease surveillance systems. **Clayton R. Fink** is a senior software engineer at APL focused on applying natural language processing and machine learning in different data domains. **Yevgeniy Elbert** is an APL statistician who contributes to data analysis and method implementation and evaluation for electronic disease surveillance tools developed at APL. **In-Kyu Yoon** is a physician and virologist and currently Chief of the Department of Virology at the U.S. Armed Forces Research Institute of Medical Sciences. His colleagues from the Philippines AFRIMS Virology Research Unit, **John Mark Velasco** and **Agnes D. Tomayao**, coordinated the field aspects of the project. Three additional medical epidemiologists from the Republic of Philippines were essential to this study for arranging access to critical data sets and providing perspective on the challenges of disease surveillance in the Republic of Philippines. They are **Vito Roque Jr.,** Head of Public Health Surveillance and Informatics Division of the National Epidemiology Center; **Enrique Tayag**, Assistant Secretary, Public Health Surveillance and Informatics Division of the National Epidemiology Center; and **Durinda Macasocol**, Assistant Epidemiologist at the Cebu City Health Office. **Sheri H. Lewis** is the Global Health Surveillance Program Manager in the Asymmetric Operations Sector's Homeland Protection Program Management Office and is responsible for development of business opportunities in public health. For more information on the work reported here, contact Jacqueline Coberly. Her e-mail address is jacqueline.coberly@jhuapl.edu.

The *Johns Hopkins APL Technical Digest* can be accessed electronically at **www.jhuapl.edu/techdigest**.