Using Social Media to Enhance Disease Surveillance

Jacqueline S. Coberly, PhD
Clay Fink, PhD
Eugene Elbert, MS
In-Kyu Yoon, MD
John Mark Velasco, MD, MPH
Sheri Lewis, MPH

NDIA BioSurveillance Conference
28 August 2012
Disclaimer

The views expressed in this presentation are those of the authors and do not reflect the official policy of the Department of the Army, Department of Defense, or the U.S. Government.
Outline

• Disease Surveillance Program at JHU/APL
• Can Twitter provide valid surrogate data to inform electronic disease surveillance
  – Twitter Project Objective
  – Methods & Results
• Conclusions
Electronic Syndromic Surveillance

Alert is identified for a particular day / syndrome

ED Chief Complaints

Poison Control

Rx Drugs

Nurse Call Center

School Absenteeism

Radiology

Diagnostic Labs

Ambulance Logs

Epidemiologist performs daily system review

Epidemiologist gathers additional data
• Surveillance data
• Lab reports
• Facility reports
• Verbal reports

School Absenteeism

More detailed analysis of alert

Outbreak Confirmed

PUBLIC HEALTH RESPONSE INITIATED

Alert is identified for a particular day / syndrome
Evolution of ESSENCE and Electronic Disease Surveillance at JHU/APL

- Maryland
- National Capitol Region
- United States
- Global

Funding provided by Armed Forces Health Surveillance Center, Division of GEIS Operations
JHU/APL Global Involvement

- Collaboration with Global Emerging Infection Systems (GEIS), now part of AFHSC
- Initially Assess the utility of syndromic surveillance in resource poor areas
- Currently Develop, implement, support and evaluate integrated global disease surveillance and response software system.
SAGES
(Suite for Automated Global Electronic bioSurveillance)
Twitter as a Surveillance Data Source?

• Project Objectives
  – To investigate whether Twitter data can be used to detect & characterize the incidence of dengue-like fever in a dengue-endemic area.
  – Compare Twitter ‘dengue’ trend data with ‘fever’ and dengue incidence data collected by local and national health authorities.

• Limited pilot done on internal R&D funding
Twitter Project – Methods Overview

• Obtain ‘ground truth’ data for 2011
• Collect publicly available Twitter messages during 2011 dengue season.
• Identify a vocabulary of words/phrases in tweets
• Perform keyword analysis using vocabulary; compare marked tweets with SMS-C and PIDSR
Ground Truth Data

• Two sources
  – Fever SMS data, Cebu City, PI (SMS-C)
  – Nat’l Reportable Disease System, Dengue (PIDSR)
Ground Truth Data
Fever SMS Program, Cebu City

• Fever incidence mimics dengue incidence
• Paper based fever reporting system used in Cebu City until 2009
• Replaced by city-wide fever reporting via SMS
  – Each local clinic texts data for each patient presenting with fever to the Cebu City Health Office (CCHO) daily
Ground Truth
SMS-C Data vs. Adjusted SMS-C Data

[Graph showing the comparison between SMS-C 7dAvg and SMS-C data over time from 6/18/2011 to 9/10/2011.]
Ground Truth Data
National Reportable Disease System

- Philippines Integrated Disease Surveillance and Reporting system (PIDSR)
- Each case of reportable disease observed, including dengue, is reported to the National Epidemiology Center
- Covers entire country
- Detailed case report, but not timely
Ground Truth
PIDSR - Cebu City, NCR & Combination

Pearson Correlation Coefficient = 0.594, p<0.0001
Collection of Tweets

• From 2 areas of the Philippines:
  – Cebu City (C)
  – National Capitol Region (NCR)

• Time period:

• From Twitter public Application Program Interface (API)
  – Prospective only
  – Only a fraction of total, exact method of selection is unclear
Tweets by Location: Cebu City vs. NCR

The diagram compares the number of tweets from Cebu City and the National Capital Region (NCR) over a period starting from June 18, 2011, to September 10, 2011. The x-axis represents the dates, and the y-axis shows the number of tweets. The line graph indicates fluctuations in tweet volumes with peaks and troughs across the specified period.
Keyword Analysis Results

• Dengue
  – Few mentions n=287 (~0.001%)
  – Most from public health/news announcements

• Clinical diagnosis (fever and \( \geq 1 \) other sx)
  – Increased specificity
  – Still relatively few mentions, N=441 (~0.002%)

• Fever
  – Traditionally used as a surrogate for dengue
  – Most frequent appearance, N=8814 (~0.03%)
  – Medically related fever is less common (N=4409), but more relevant
Description of Tweets

- **10,303,366** Cebu Tweets
- **15,719,767** NCR Tweets

- **9,461** Fever Not ‘Beiber’
  - **644** Duplicates

- **8,814** All Fever
  - (Cebu – 2,472)
  - (NCR – 6,746)

- **4099** Fever, Medically Related
  - (Cebu – 995)
  - (NCR – 3,104)

- **4715** Fever <> Medical

Used **combined** Cebu City & NCR tweets
All Fever vs Medically Related Fever Tweets

Number Tweets


AllTweets   TweetsMedRltd

Used medically-relevant tweets
Medically-Related Fever Tweets vs Adjusted SMS-C

Pearson Correlation Coefficient = 0.575, p<0.0001
Medically-Related Fever Tweets with 6 day Shift vs Adjusted SMS-C

Pearson Correlation Coefficient = 0.769, p<0.0001
Medically–Related Fever Tweets vs Nat’l Reportable Disease (PIDSR-C&NCR)

Pearson Correlation Coefficient = 0.629, p<0.0001
Medically-Related Fever Tweets w/12d Shift vs Nat’l Reportable Dengue (PIDSR-C&NCR)

Pearson Correlation Coefficient = 0.829, p<0.0001
Limitations

- Limitations on ‘free’ tweets from Twitter
- Issues with the Twitter data feed necessitated combining tweets from Cebu City and NCR
Conclusions

• Twitter leads SMS-C data by 6 days and PIDSR-C&NCR data by 12 days

• This suggests that Twitter data may be a useful and timely source of data for automated disease surveillance

• Further investigation is needed
  – Repetition with 2012 data to resolve data collection errors
  – More sophisticated machine learning techniques
  – Implementation into an electronic surveillance system
<table>
<thead>
<tr>
<th>JHU/APL</th>
<th>AFRIMS</th>
<th>NEC</th>
<th>CCHO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheri Lewis</td>
<td>In-Kyu Yoon</td>
<td>Enrique Roque</td>
<td>Ilya A. Tac-an Abellanosa</td>
</tr>
<tr>
<td>Jacqueline Coberly</td>
<td>John Mark Velasco</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brian Feighner</td>
<td>Maria Theresa Alera</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rekha Holtry</td>
<td>Agnes Tomayo</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vivian Hung</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Richard Wojcik</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Timothy Campbell</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjoa Poku</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Charles Hodanics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Howard Burkom</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Related Research


