The Complex Relationship of Real-space Events and Messages in Cyberspace: A case study of Influenza and Pertussis using Tweets

Authors: Anna C. Nagel, Ming-Hsiang Tsou, Li An, Jean Marc Gawron, Dipak K Gupta, Brian Spitzberg, Jiue-An Yang, Su Han, K. Michael Peddecord, Mark H. Sawyer, Suzanne Lindsay

Abstract

Background - Syndromic surveillance plays a vital role in disease detection, but the traditional methods of collecting patient data, reporting to health officials, and then compiling a report are costly and time consuming. In recent years syndromic surveillance tools have expanded and researchers are able to exploit the vast amount of data available in real time on the internet at almost no cost. While many data sources for *infoveillance* exist, in this study we focus on status updates (tweets) from the popular microblogging website Twitter.

Objective - In this study we aim to explore the interaction between cyberspace, measured by tweets, and real world influenza and pertussis occurrence. Tweets were aggregated by week and compared to weekly influenza-like illness (ILI) and weekly pertussis incidence. In addition, tweets were subdivided into four categories; non re-tweets, re-tweets, tweets with a URL web address, and tweets without a URL web address to investigate which groups of tweets correlated best with disease incidence.

Methods - Tweets were collected within a 17 mile radius of 31 US cities based on population size. Influenza (flu) analysis was restricted to the 11 cities with sufficient ILI data. Pertussis analysis was based on the two cities nearest to the Washington State pertussis outbreak; Seattle, WA and Portland, OR. Tweet collection resulted in 459,043 "flu," 16,761 "influenza," 2,823 "pertussis," and 16,690 "whooping cough" tweets. The correlation coefficients between tweets or subgroups of tweets and disease occurrence were calculated and trends were represented in barcharts for easy visualization.

Results - Correlations between weekly aggregated tweets and disease occurrence varied greatly, but were relatively high in some areas. In general correlation coefficients were higher in the flu analysis compared to the pertussis analysis and within each analysis "flu" tweets were better correlated with ILI than "influenza" tweets and "whooping cough" tweets correlated better with pertussis incidence than "pertussis" tweets. In general non re-tweets correlated better with disease occurrence than re-tweets and tweets without a URL web address better than those with a URL web address, however these comparisons were mostly just significant for the "flu" tweets.

Conclusions - This study demonstrates that not only does keyword choice play an important role in how well tweets correlated with disease occurrence, but the subgroup of tweets used for analysis was also important. Our exploratory work shows potential in the use of tweets for *Infoveillance*, but continued efforts are needed to further refine research methods in this field.

Keywords - Twitter, tweets, infoveillance, infodemiology, cyberspace, syndromic surveillance, influenza, pertussis/whooping cough.