Exploring the Metadata of Vaccine-Related Twitter Posts: Just How Much Activity Is There and Where Does It Come from?

Jason Brinkley, PhD[1]; Sarah W. Ball, ScD, MPH[1]; Alison Thaung, MBA[1]; Zoran Obradovic, PhD[2]; Marija Stanojevic[2]; Fang Zhou[2]; Stacie M. Greby[2]; Cindi Knighton[2]; Allison Fisher[2]

[1]Abt Associates; [2]Temple University; [3]Immunization Services Division, National Center for Immunization and Respiratory Diseases, CDC

Background

- Twitter has become a popular platform for the dissemination of health information, advice, and opinions.
- Researchers are looking to measure the frequency, scope, reach, and sentiment of tweets based on key words found within the tweets themselves.
- Language processing and text analytics have become an instrumental part in that work.
- There is a wealth of secondary metadata generated from Twitter activity that goes beyond the content of the tweets themselves.
- As part of an ongoing Centers for Disease Control and Prevention (CDC) Immunization Program Cognitive Computing System (IPCCS) Pilot, immunization related Twitter data was collected over a 15 month period.
- The development of a usable search lexicon of terms related to vaccination was the primary goal of the project, and that lexicon was built with support from a variety of sources.
- As an add-on project, we looked at the meta-data of those who posted tweets with certain vaccine related key words.

Methods

- The Sysomos Search tool was used to help search and pull data off of the web for lexicon development. Search terms included 'vaccination', 'cancer', 'hpv', 'hepatitis', pneumococcal', 'shingles', 'measles', and others along with variations, abbreviations, and derivatives .
- Only Twitter accounts identified as based in the United States, including government and non-government sponsored accounts, were included.
- Collected tweets originated from July 31, 2016 to October 29, 2017.
- Across the 15 month time period, we scraped 3,764,278 unique tweets (and retweets) available for analysis.
- There were 1,022,594 unique Twitter accounts who posted tweets with keywords across the analyzed time period averaging of 3.7 tweets per account.

Results

Exhibit 1. Unique Authors and Tweet Counts

Tweet Frequency Group	Number of Keyword Accounts	Percent of Total Keyword Accounts	Number of Tweets	Percent of Total Tweets
More than 500 Tweets	275	0.03%	365,373	9.7%
10 to 500 Tweets	46,243	4.52%	1,698,762	45.1%
Two to Nine Tweets	324,161	31.70%	1,048,228	27.9%
One Tweet	651,915	63.75%	651,915	17.3%
All Sources	1,022,594	100%	3,764,278	100%

- While there is a lot of activity using these keywords, much of the discussion was being driven by a small number of accounts: 55% of all Twitter activity using vaccination-related terms was generated by approximately 5% of the Twitter accounts that used any of the keywords
- None of these *high frequency accounts* were linked to an official United States federal agency. The majority of Twitter accounts (64%) that used any of the search keywords used them only once.

Exhibit 2. Measuring Reach - Number of Followers (Log Scale) by Number of Tweets (Log Scale)

- We found that among the 5% high frequency accounts which generated the most vaccination-related activity, the median number of followers was around 1,000 users.
- By comparison the official CDC Twitter account, @CDCgov, posted vaccination-related tweets about half as frequently as these accounts and has over 845 thousand followers (at time of analysis).

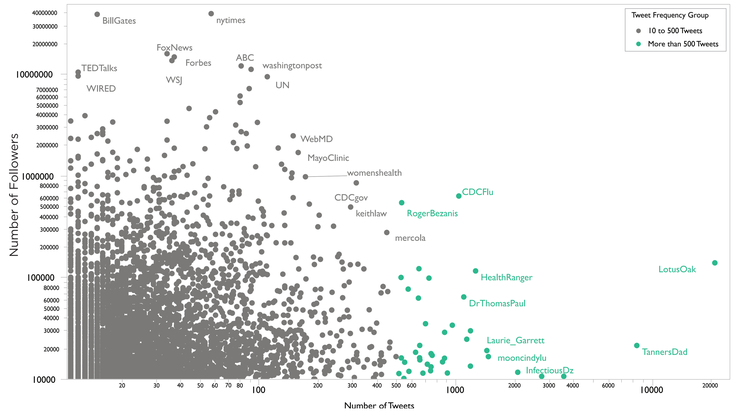
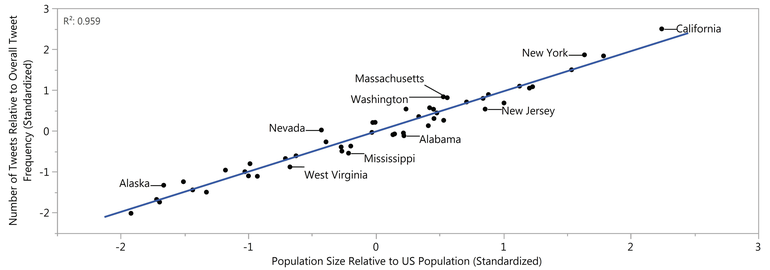


Exhibit 3. Standardized Tweet Frequency per State Plotted by Population Size

- We standardize both tweet frequency within state and population size (from Census) metrics and plot state size against tweet frequency, which we have done below in Exhibit 3.
- The line represents the expected tweet frequency for a particular state's size (standardized).
- States with high deviations are highlighted, so we can see that California has slightly more Twitter activity in this area than their population size suggests. However, relative to their population size, they post no more frequently than, for example, Alaska or Nevada.



Conclusion

We conclude that Twitter metadata can be a rich and informative source of information about the size, time, and locations of discussions on specific topics. There is a great limitation in not knowing the content of those discussions and how keywords are being used (i.e. sentiment).

Implication for Policy or Practice

More work needs to be done and better tools developed to help understand Twitter content, but our work shows that the metadata can be an important starting point for understanding the size and scope of social media activity.

For more information, please contact
Jason Brinkley, PhD
Jason_Brinkley@abtassoc.com

Acknowledgements
This project is supported under HHS/CDC contract HHSD2002014M60311B, 2018.

Disclaimer
The authors have no conflicts of interest to disclose.