SCHOOL OF
PUBLIC HEALTH
CENTER FOR HEALTH EQUITY

BOTH MIRROR AND COMPLEMENT: A COMPARISON OF SOCIAL MEDIA DATA AND SURVEY DATA ABOUT FLU VACCINATION

David Broniatowski, Ph.D., George Washington University; Mark Dredze, Ph.D., Johns Hopkins University; Karen Hilyard, Ph.D., University of Georgia; Maeghan Dessecker, M.A., M.P.H., University of Georgia; Amelia Jamison, M.A., M.P.H., University of Maryland; Michael Paul, Ph.D., University of Colorado, Boulder; Michael Smith, M.S., George Washington University; Sandra Crouse Quinn, Ph.D., University of Maryland

BACKGROUND

Social media offers researchers opportunities to assess public knowledge, attitudes, and health behaviors in real time. Researchers have relied on survey methods that provide reliable and accurate data, but are time consuming. Survey research struggles to capture the attitudes of young people, minority groups, and urbanites - all groups well represented on social media. However, social media brings unique challenges including huge amounts of data and difficulty in accurately assessing demographic information about social media users.

BACKGROUND

The goal of this study is to compare Twitter data to published CDC survey data on influenza vaccination, documenting the ways that the data complement each other, where they diverge, and importantly, how these patterns may differ across demographic groups.

RESEARCH QUESTIONS

Using natural language classifiers to infer demographic information and vaccine intentions from Twitter, we compared these findings to published CDC data.

Can Twitter data to be used to detect the following:

- Changes in vaccine uptake during the flu season?
- Changes in vaccine uptake between flu seasons?
- Differences in influenza vaccination by gender?
- Differences in vaccination rates between geographic locations?

MEASURES & METHODS

TRADITIONAL SURVEY DATA

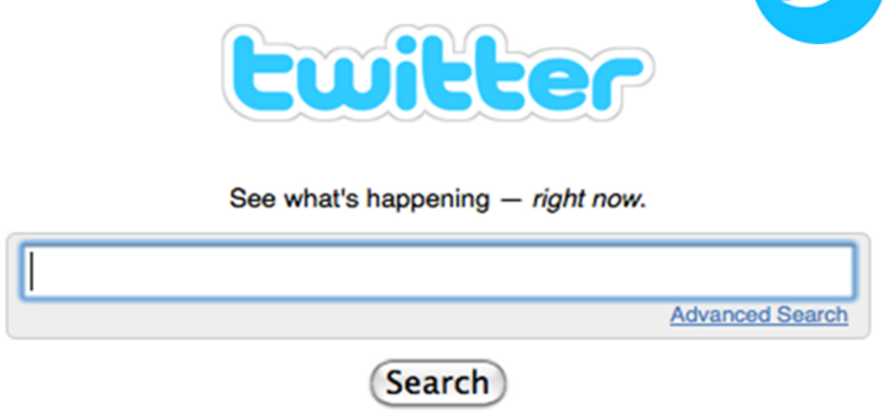
- CDC's "FluVaxView"1: data from nationally representative surveys including the National Immunization Survey (NIS-Flu), the Behavioral Risk Factors Surveillance System (BRFSS) and the National Health Interview Survey (NHIS).
- Vaccination rates are estimated by age, gender, geographic location (HHS Regions) and for each month during the flu season.

FluVaxView

MEASURES & METHODS

TWITTER DATA

- Since 2012, collected tweets that contain health-related keywords, using the Twitter search API.2
- A subset of "flu-related tweets" were identified as having one flu-related keyword (flu, influenza) and one vaccine-related keyword (shot, vaccine, vaccination). Retweets and Non-English tweets were excluded.



MEASURES & METHODS

TWEET CLASSIFICATION

- Data: 1,007,582 flu-related tweets
- Used Amazon Mechanical Turk to annotate a random sample of 10,000 tweets. Annotators were asked the following:
  - Does this message indicate that someone received, or intended to receive, a flu vaccine? (Y/N)
  - If Yes, clarify vaccine intent or receipt of vaccine.
- Collected 3 independent annotations per tweet

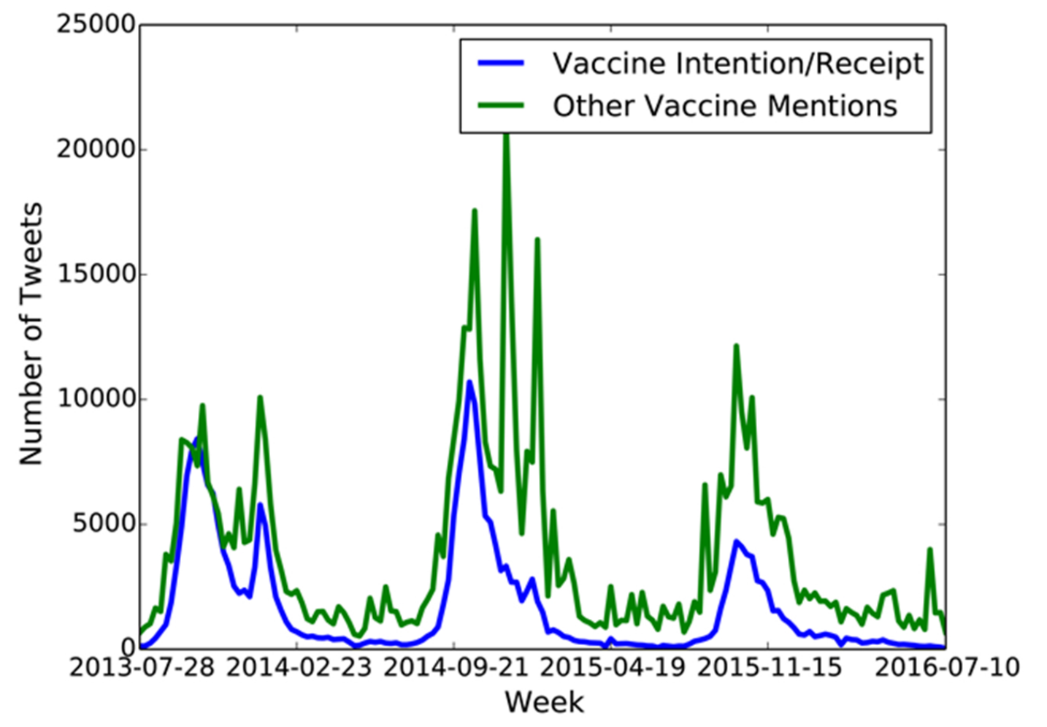
amazon mechanicalturk
Artificial Artificial Intelligence

MEASURES & METHODS

TWEET CLASSIFICATION (CONT.)

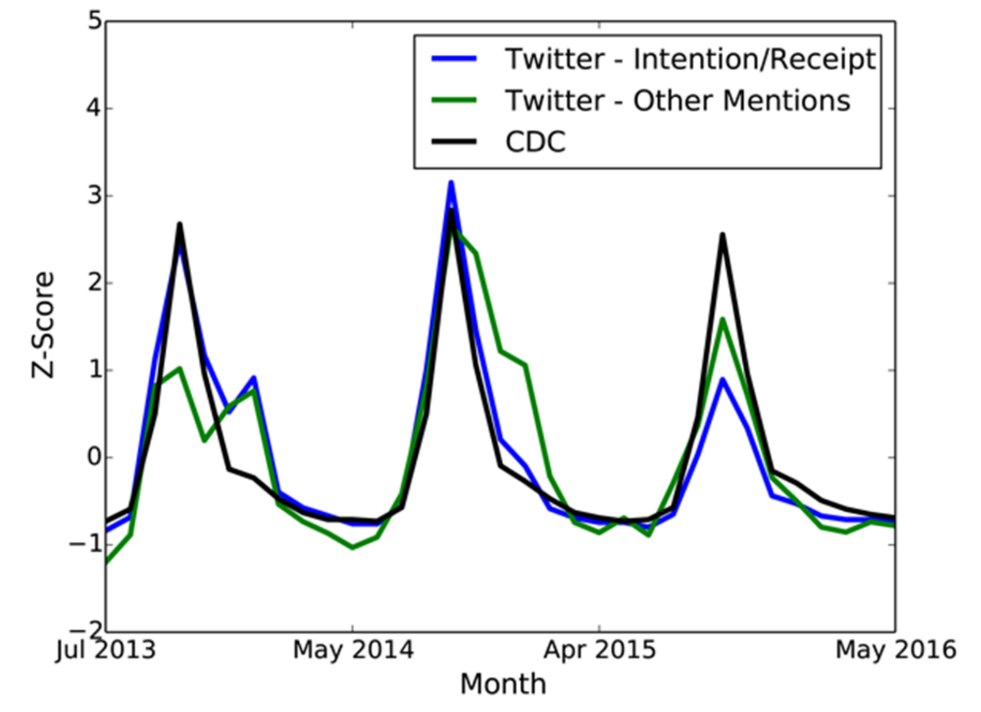
- We fit a logistic regression model to the data.
- We converted each tweet by counting the frequency of each word (or frequently-occurring groups) in each tweet and weighting these frequencies by a measure of how frequently that word appears across all tweets. Emojis ☺ and emoticons :) were treated as if they were words.
- We used an algorithm developed by Gimpel et al3 to infer frequently-occurring linguistic syntax constructions, which were treated as words.
- Approximately 30.5% of the tweets were identified as indicating vaccine intention/receipt, and of that set, 88% indicated that vaccine receipt

RESULTS: Weekly Counts of Vaccine Tweets



- The smoothness of the blue line compared to the green line suggests our classifier is reducing the noise surrounding vaccine tweets.

RESULTS: Vaccine Uptake on Twitter Compared to CDC Data

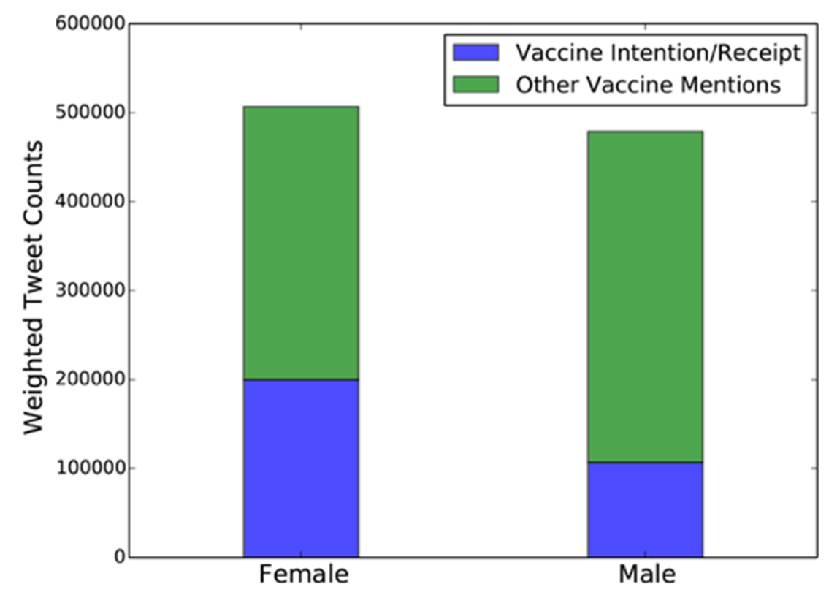


- We utilized standardized counts (z-scores) to make the Twitter and CDC data comparable.

RESULTS: Vaccine Uptake on Twitter Compared to CDC Data (CONT.)

- The correlation is stronger between the vaccine intention/receipt tweets and the CDC data (r=.903) than the between the general vaccine tweets and the CDC data (r=.816). This was not statistically significant.
- Analysis revealed a stronger correlation between vaccine receipt tweets and the CDC data (r=.911) than between vaccine intention tweets and the CDC data (r=.869). This was not statistically significant.

RESULTS: Vaccine Uptake by Gender



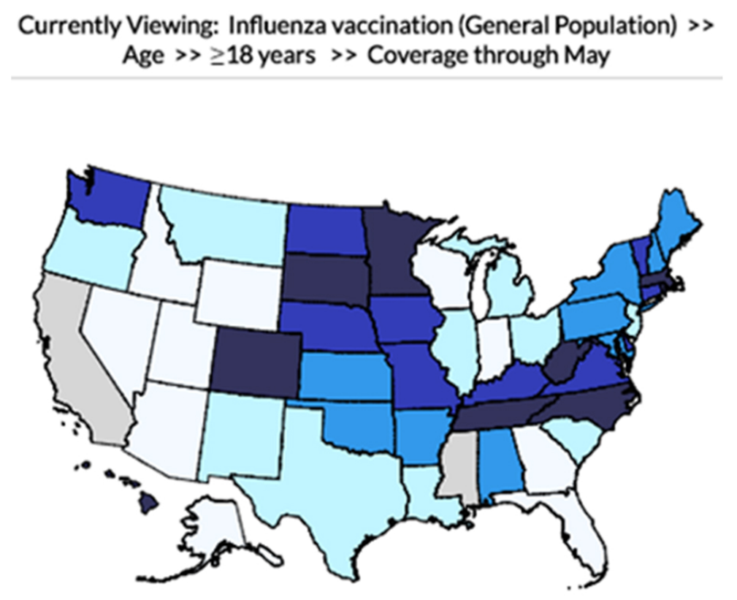
Inferred gender using the Demographer Tool5 and counts were adjusting for uneven Twitter use.

Roughly the same number of tweets by gender, with slightly more tweets by females. Females are significantly more likely (p<.01) to tweet about vaccine intentions/receipt.

During the 2015-2016 flu season, 38% of adult males were vaccinated compared to 45% of adult females.[1]

RESULTS: Vaccine Uptake by Geography

CDC Estimates 2013-2014


Currently Viewing: Influenza vaccination (General Population) >>
Age >> ≥18 years >> Coverage through May

Using the 2013-14 Flu Season as a test case, we aggregated tweets into 10 HHS regions.

Geographic location was inferred using Carmen geolocation system,4 and normalized location-specific counts.

Result: strong correlation between the vaccine intention/receipt tweets and the CDC's regional vaccine uptake percentages (r=.711).

DIRECTIONS FOR FUTURE RESEARCH

Develop methods to infer additional demographics including age and race/ethnicity.

- Move beyond validating existing CDC statistics to collect data not already captured in existing research

Improving classifiers through extensive parameter tuning and better features.

- Building a classifier to effectively capture vaccine sentiment in addition to vaccine intentions.

CONCLUSIONS

- Twitter data is a promising means of tracking flu vaccine behavior.
- Analysis of these data require new tools in order to be more useful to public health professionals.

ACKNOWLEDGEMENTS

Supplementing Survey-Based Analyses of Group Vaccination Narratives and Behaviors Using Social Media, David Broniatowski & Sandra Quinn, PI, Grant 5R01GM114771, funded by the National Institute of General Medical Sciences, NIH.

CITATIONS

1. Centers for Disease Control and Prevention (2016). Flu Vax View: Interactive. http://www.cdc.gov/flu/fluvaxview/
2. Paul, M. J., and Dredze, M. (2014) Discovering health topics in social media using topic models. PLoS ONE 9(8):e103408.
3. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., et al. (2011). Part-of-speech tagging for Twitter: Annotation, features, and experiments. In Association for Computational Linguistics (ACL).
4.Dredze, M.,Paul, M. J., Bergsma, S., & Tran, H.(2013) Carmen:A Twitter geolocation system with applications to public health. In AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI).
5. Knowles, R., Carroll, J., & Dredze, M. (2016) Demographer:Extremely simple name demographics. In EMNLP Workshop on Natural Language Processing and Computational Social Science