Title:

Can Online Self-Reports Assist in Real-Time Identification of Influenza Vaccination Uptake? A Cross-Sectional Study of Influenza Vaccine-Related Tweets in the US, 2013-2017

Authors:

Xiaolei Huang[1], Michael C. Smith[2], Amelia M. Jamison[3], David A. Broniatowski[2], Mark Dredze[4], Sandra C. Quinn[3,5], Justin Cai[6], Michael J. Paul[1,6]

[1] Department of Information Science, University of Colorado, Boulder, CO 80309, USA
[2] Department of Engineering Management & Systems Engineering, George Washington University, Washington, DC 20052, USA
[3] Center for Health Equity, School of Public Health, University of Maryland, College Park, MD 20742, USA
[4] Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA
[5] Department of Family Science, School of Public Health, University of Maryland, College Park, MD 20742, USA
[6] Department of Computer Science, University of Colorado, Boulder, CO 80309, USA

Corresponding Author:

Michael J. Paul
Assistant Professor, Department of Information Science
315 UCB, Boulder, CO 80309, USA
1-217-552-3605
mpaul@colorado.edu

Word Count: 3,303

ABSTRACT

Introduction: The Centers for Disease Control and Prevention (CDC) spend significant time and resources to track influenza (flu) vaccination coverage each flu season using national surveys. Emerging data from social media provide an alternative solution to surveillance at both national and local levels of flu vaccination coverage in near real-time.

Objectives: This study aimed to characterize and analyze the vaccinated population from temporal, demographic, and geographical perspectives using automatic classification of vaccination-related Twitter data.

Methods: In this cross-sectional study, we continuously collected tweets containing both flu-related terms and vaccine-related terms covering four consecutive flu seasons from 2013 to 2017.We created a machine learning classifier to identify relevant tweets, then evaluated the approach by comparing to data from the CDC's FluVaxView. We limited our analysis to tweets geolocated within the US.

Results: We assessed 1,124,839 tweets. We found strong correlations of .799 between monthly Twitter estimates and CDC, with correlations as high as .950 in individual flu seasons. We also found that our approach obtained geographic correlations of .387 at the US state level and .467 at the regional level. Finally, we found a higher level of flu vaccine tweets among female users than male users, also consistent with the results of CDC surveys on vaccine uptake.

Conclusion: Significant correlations between Twitter data and CDC data show the potential of using social media for vaccination surveillance. Temporal variability is captured better than geographic and demographic variability. We discuss potential paths forward for leveraging this approach.

Keywords: vaccination, surveillance, influenza, biostatistics, time-series

ARTICLE SUMMARY

Strengths and limitations of this study

- This study shows how to measure influenza vaccination uptake through Twitter, which has advantages and disadvantages compared to traditional survey methods.
- The signal from Twitter is available in real-time and has potential to be localized to specific geographic locations.
- While Twitter can be considered “big data”, the sample size is more limited when narrowed to specific populations.
- Certain vulnerable populations, including children and older adults, are underrepresented in Twitter data.

INTRODUCTION

The Advisory Council for Immunization Practices (ACIP) at the Centers for Disease Control and Prevention (CDC) recommends annual influenza vaccination for all healthy adults.[1] Furthermore, CDC urges individuals to get vaccinated early in the flu season, from October through January.[2] Yet, it can be difficult for researchers and practitioners working to improve influenza vaccine uptake to get accurate information in real time. Existing influenza immunization surveillance techniques have known limitations: traditional survey-based methods are time-consuming and expensive, and newer reimbursement-based systems fail to accurately capture a representative sample of population.[3]

Two national surveillance systems enable public health professionals to access information on influenza vaccine uptake in the United States (US). The most accessible of these systems is the CDC's FluVaxView, which aggregates uptake data from several national surveys.[4] The CDC data provide accurate estimates of vaccine uptake, although with some time lag. The earliest reports are only available after flu seasons typically peak, and final estimates are generally published at the start of the following flu season in September or October. Additionally, the panel surveys that inform the reports are expensive, take months to administer and process, and may undersample populations without a landline phone, particularly minority populations, young adults, and adults living in urban areas.[5, 6] A second system,[7] provided by the National Vaccine Program Office, uses an online tool to "live-track" influenza vaccination insurance claims from Medicare beneficiaries. While this system reduces lag time between vaccination and reporting, it only captures the population enrolled in Medicare, adults over age 65 and those under 65 living with disabilities.[7]

Social media data have been utilized in new tools for infectious disease surveillance, particularly for seasonal and pandemic influenza.[8-10] Utilizing data from social media platforms (like Twitter or Facebook), search engines (like Google), and other internet-based resources (like blogs), researchers have been able to track the spread of disease in real time with relatively high accuracy.[9] A recent meta-analysis of social media influenza surveillance efforts found that in a comparison to national health statistics (primarily from the CDC), correlation between social media data and national statistics ranged from 0.55 to 0.95,[11, 12] and the majority of projects were able to predict outbreaks more quickly than traditional surveillance methods.[10] Of these studies, the most accurate systems have harnessed natural language processing methods to identify relevant tweets. However, few of these tools have been fully integrated into public health practice.

With the development of new tools and techniques, social media data have the potential to similarly inform the practice of influenza immunization surveillance. However, to our knowledge, no studies have attempted to utilize social media data to track influenza vaccine intentions and uptake at the national level. To date, efforts to track influenza vaccination through social media have been much less frequent than efforts to track disease. Researchers are more likely to focus on the use of social media as a health communication tool than to explore the potential for immunization surveillance.[13] Some studies have been able to use social media data to track vaccine sentiment and general attitudes towards vaccines.[14–16] Others have focused on the spread of vaccine sentiment across online social networks.[17, 18] Some vaccine-specific studies have also attempted to use social media to identify geographic differences in vaccine uptake.[19, 20] The possibility of efficiently tracking influenza immunization in real-time is promising, but the true value of any new data source is limited without validation against known metrics.[14, 21, 22] To successfully use social media data in immunization surveillance efforts, an important first step is to validate observed trends against national survey data. In this study, we sought to validate observed patterns from Twitter, using tweets expressing either intention to seek immunization or receipt of influenza immunization, against influenza immunization data from the CDC for four consecutive flu seasons from 2013-2017.

METHODS

Patient and Public Involvement

This study did not involve patients. This work was conducted under Johns Hopkins University Homewood IRB No. 2011123: "Mining Information from Social Media", which qualified for an exemption under category 4.

Data

Twitter Data

We continuously collected tweets containing the terms “flu” or “influenza” since 2012 using the Twitter streaming Application Programming Interface (API), as part of data described in our team’s prior work on Twitter-based health surveillance.[23] For this study, we filtered influenza-related tweets containing at least one vaccine-related term (“shot(s)”, “vaccine(s)”, and “vaccination”). We then inferred the US state for tweets using the Carmen geolocation system,[24] and the gender of each Twitter user of the dataset using the Demographer tool.[25] The Carmen tool infers locations of tweets by three main sources, coordinates of tweets, places name of tweets and locations in user profiles, and most often represents the home location of the user rather than their location while tweeting. The Demographer tool infers binary genders of Twitter users by the names of their profiles. We removed retweets, non-English tweets and tweets not located in the US. We obtained 1,124,839 tweets from 742,802 Twitter users covering four consecutive flu seasons from 2013 to 2017. More details can be found in the supplementary material (A1 and A2).

In addition to tweets about influenza vaccination, we also collected a random sample of tweets from all of Twitter. This was used to adjust the vaccine counts by time, location, and demographics, as described below. The random sample includes approximately 4 million tweets per day since 2011.

CDC Data

We utilized CDC data on influenza vaccination of the four flu seasons for validating our approaches. The CDC data were downloaded from the CDC's FluVaxView system.[4] These data include vaccination coverage by month, by states, and by geographic regions as defined by the US Department of Health and Human Services (HHS). The CDC's estimates are based on several national surveys: the Behavioral Risk Factor Surveillance System (BRFSS, which targets adults), the National Health Interview Survey (NHIS), and the National Immunization Surveys (NIS, which focuses on children). In this study, we use the CDC data for adults (≥18 years old) across all racial/ethnic groups. The CDC reports the "sex" of the respondents, although the underlying surveys ask for "gender" rather than sex,[26, 27] making this variable comparable to our definition of gender in Twitter.

Automated Classification

In our study, we used natural language processing techniques to preprocess and encode tweets into feature vectors, then used the vectors to build machine learning classifiers to automatically categorize Twitter messages that express vaccination behavior. Tweets were classified into yes or no labels in response to the question, "Does this message indicate that someone received, or intended to receive, a flu vaccine?" Specifically, we randomly sampled 10,000 tweets from our collected data from 2012 to 2016 and then used a crowdsourcing platform to annotate the 10,000 tweets,[28] using quality control measures to ensure accurate annotations. The classifiers were trained by the annotated tweets.

The best-performing classification model was a convolutional neural network (CNN), which had a precision (the proportion of tweets classified as vaccine intention/receipt that were correctly classified) of 89.4% and recall (the proportion of vaccine intention/receipt tweets that were identified by the classifier) of 80.0%, measured using nested five-fold cross-validation. This classifier was applied to the full dataset of 1,124,839 tweets, of which 366,698 were classified as expressing that someone received or intended to receive an influenza vaccine. More details of preprocessing and encoding tweets, and building and selecting machine learning models, can be found in the supplementary materials (A.2) as well as in our prior preliminary work using simpler models.[29]

Trend Extraction and Validation

To evaluate the reliability of the Twitter classification model as a source for vaccination surveillance, we compared the Twitter data to CDC data along three dimensions: time (by month), location (by US state and region), and demographics (by gender). Specifically, CDC FluVaxView provides the monthly percentage of American adults who received an influenza vaccination in a given month in each state, as well as the percentage of Americans who report vaccination in different demographic groups each flu season.

To extract trends over time, we computed the number of vaccine intention/receipt tweets in each month per season, excluding June (the CDC does not report data for June). We only included tweets geolocated to the US. To adjust for variations in Twitter over time, we divided the monthly counts by the number of tweets in the same month from the large random sample of tweets.[8] In addition to monthly rates for direct comparison to CDC, we also calculated weekly tweet rates, providing estimates at a finer time granularity than reported by the CDC. For monthly time series data, we applied an autoregressive integrated moving average (ARIMA) model and linear regression to estimate the CDC data from the Twitter data.[30]

To extract trends by location, we computed the number of intention/receipt tweets in each of the 10 HHS regions and each of the 50 US states. We created per-capita estimates by dividing each count by the number of tweets from the same region or state from the random sample of tweets.

To extract trends by gender, we computed the number of intention/receipt tweets identified as male or female, divided by the corresponding counts from the random sample. We computed this proportion within each US state before aggregating the counts from all states, to additionally adjust for gender variation across location. We provided detailed validation steps and additional experiments in supplementary material A.3.

Confidence Intervals

We present 95% confidence intervals for all results. There are two sources of variability we must account for when constructing confidence intervals. One source is the set of points included in the correlation. The other is the set of tweets used to estimate the level of vaccine intention in each group. When estimating values within fine-grained groups, such as specific US states, the number of tweets can be small, leading to high variability in the estimates that propagates to the estimate of the correlation.

To address these issues, we construct confidence intervals using bootstrap resampling.[31] We perform sampling at two levels. First, we sample the set of tweets used to calculate the estimate in each group (e.g., the tweets in a specific month or location). We then sample the set of points that are included in the calculation of the correlation (e.g., the set of months). The confidence intervals are constructed from 100 bootstrap samples.

RESULTS

Activity by Time

Table 1 shows the correlation between the classified tweets and CDC data from the ARIMA results along with 95% confidence intervals. Figure 1 shows the values from both data sources over time, standardized with z-scores. While the CDC data are only available by month, we show Twitter counts by week (Sunday to Saturday) to illustrate the finer temporal granularity that is possible. In both data sets, there are seasonal peaks every October, when influenza vaccines are distributed in the US. While the overall shapes are very similar, the Twitter data sometimes shows rises later in the flu season that do not correspond to a similar rise in the CDC data, especially in the 2013-14 season, which results in the lowest correlation.

Table 1. Pearson correlations (95% CI) by month in each flu season.

	All seasons	2013-14	2014-15	2015-16	2016-17
Monthly	.799 (.797 - .801)	.644 (.639 - .647)	.950 (.948 - .951)	.909 (.905 - .913)	.910 (.909 - .912)

Activity by Location

The prevalence of tweets mentioning vaccine intention/receipt in each location is shown in Figure 2, where darker color indicates more frequent vaccine mentions. We observe that states in the northwest, especially Washington and Oregon, have higher rates than southeastern states, such as Florida and Alabama. There is a moderate correlation between the geographic patterns in the Twitter data compared to the CDC data, with a higher correlation at the HHS region level than at the state level (Table 2). The strength of the correlations varies by season, with much stronger correlations in the first two seasons than the latter two seasons.

Table 2. Pearson correlations (95% CI) by geography in each season.

	All seasons	2013-14	2014-15	2015-16	2016-17
State	.387 (.362 - .394)	.300 (.261 - .308)	.214 (.193 - .243)	.051 (.015 - .057)	.025 (.002 - .040)
HHS Region	.467 (.445 - .483)	.690 (.650 - .714)	.573 (.539 - .600)	.137 (.090 - .179)	.244 (.213 - .272)

Activity by Gender

Female users are much more likely to tweet about vaccine intention/receipt than male users on Twitter. The female-to-male ratios in each of the four seasons are (with 95% CIs), respectively: 1.97 (1.96 - 1.98), 1.73 (1.72 - 1.74), 1.59 (1.58 - 1.59), 1.47 (1.46 - 1.48). This ratio is higher than in the CDC data (1.18, 1.17, 1.19, 1.20). However, the two data sources are in relative agreement: the vaccination rate is higher among females than males. For example, in the 2016-17 flu season, the CDC reported that among American adults, 47.0% of women were vaccinated for influenza, compared to 39.3% of men.

We visualized the gender weekly trends and gender ratio of vaccine coverage across locations in Figure 3. The plot of gender weekly trends shows the volume of vaccine intention/receipt tweets over time. The gender ratio has also decreased steadily over time in the Twitter data, while it has stayed fairly constant in the CDC data. The plot of gender ratio shows the female-to-male ratio of vaccine intention/receipt tweets within each US state, with darker color indicating a higher ratio. For example, the figure shows that West Virginia has more females mentioning influenza vaccine behavior than males. We provided additional analyses in the supplementary material A.4.

DISCUSSION

By utilizing natural language processing techniques, Twitter data can be effectively analyzed to identify meaningful information about influenza vaccination intentions and behaviors at the population level. Our key finding is the strong correlation between monthly Twitter-based estimates of vaccination uptake and official CDC uptake estimates. Additionally, exploratory analysis suggests that natural language processing tools can be developed to further investigate significant patterns in self-reported vaccine uptake by time, location, and demographics.

Traditionally, surveillance efforts have focused on monthly or yearly data. Twitter data allows for greater flexibility and specificity when assessing temporal trends in vaccination. For example, this study shows that it is possible to extract weekly data in addition to monthly estimates. Although we are unable to compare our weekly counts to a validated national metric, we observed high week-to-week variability in general flu vaccine tweets before applying a classifier to filter out irrelevant tweets, but a relatively consistent and predictable pattern in week-to-week tweets indicating vaccine intention and receipt, suggesting that the classifiers are reducing noise at this granularity.

It is possible to capture geographic variability in Twitter data using the Carmen tool. Our results suggest some similarities with the CDC FluVaxView maps, but the associations are not strong enough to make definitive conclusions based on geography. There may be local level trends that contribute to these observed patterns. While the value of this information is limited, it does demonstrate the potential for more detailed geographic analysis in the future, especially as the number of Twitter users continues to climb.

Demographic classifiers are still under development. We were able to utilize the Demographer tool to identify the gender of the person tweeting. Our results suggest there are significantly more tweets indicating intention to vaccinate coming from females. CDC data suggest that this may be accurate, with significantly more females reporting vaccination than males according to FluVaxView. However, the gender gap in Twitter narrowed over the course of the four seasons in our study period, despite staying constant according to the CDC. Other important demographic attributes, like age, are challenging to classify and therefore not considered in this study.[32] Further refinement of demographic classifiers is necessary.

There are limitations to working with social media data. While social media is considered “big data,” we nevertheless ran into challenges with sample size. While the full dataset is indeed large, with over one million tweets, only 33.8% of those tweets can be resolved to the United States, and each experiment further filters down the data into smaller groups. For example, if tweets are counted by month within each US state, then the data needs to be split into 600 partitions (12 months times 50 states) within each year. This has an observable effect of the validity of the results: the correlations between Twitter and CDC are very strong at the national level, but weaker at the regional level, and weaker still at the state level. Sample size of tweets may also explain why the geographic correlations between Twitter and CDC (Table 2) were strong in 2013-14 and 2014-15 than in 2015-16 and 2016-17: the first two seasons contain 25.8% more geolocated tweets than the latter two seasons.

Errors in the natural language classifiers also limit overall accuracy of the approach. We investigated why the correlation with CDC was substantially lower in the 2013-14 season compared to others, and while there is no single conclusive explanation, we observed that the classifiers mis-identified flu-related tweets as indicating vaccine intentions during the peak of the flu season in January 2014, such as tweets expressing regret about not being vaccinated. This type of error was common during this month, resulting in an spike in classified tweets that did not correspond with a true rise in vaccine uptake.

These data limitations affect all social media focused research. However, among studies that utilize natural language processes to study social media data, this is one of the first studies to track vaccination uptake. Our focus on messages that explicitly indicated intention or receipt of vaccination was unique. Existing research has focused on vaccine attitudes or sentiments alone, or substitutes other measures as a proxy for behavior.[33] For example, Salanthe & Khandelwal's 2011 assessment of vaccine-related Tweets during the H1N1 influenza pandemic found strong correlation between vaccine sentiment expressed in tweets and CDC vaccine uptake rates.[17] Another study by Dunn et al. mapped exposure to negative information about HPV vaccines on Twitter to state-level vaccine uptake rates.[20] A more recent study from Tangherlini et al. focused on instances of parents opting-out of immunizations by identifying narratives describing vaccine exemptions on "Mommy blogs".[34]

Our results suggest that self-report data from Twitter can enrich the practice of influenza immunization surveillance and inform influenza vaccination campaigns. To date, the majority of social media surveillance research has been conducted without the involvement of local, state, or governmental agencies.[10] Indeed, most efforts to include public health practitioners in social media research have focused on health communications efforts.[35, 36] By utilizing an adaptable machine learning technique, research questions can be tailored to suit the needs of specific projects or organizations. For example, while we focused on estimating vaccination coverage from FluVaxView, future work could use this data in a study design that is focused on supporting decision making.[37] It may also be possible to utilize social media to track the impact and effectiveness of vaccines in a community, as early work suggests.[38]

Development of demographic classifiers for factors such as age and race/ethnicity is an important next step. One advantage of utilizing Twitter is the ability to capture behaviors from a broader range of adults, especially from groups that may be difficult to reach using traditional surveys, including young adults and members of minority groups such as African Americans and Hispanics.[30, 31] While all groups fail to reach the Healthy People 2020 recommendation of 70% uptake, these same populations (young adults and racial/ethnic minorities) are also the least likely to be immunized against seasonal influenza.[39 - 41]

Incorporating self-report social media data may allow researchers and practitioners to respond to emerging health issues in new and innovative ways, but the progress depends on the ability to integrate novel methods into existing frameworks and to validate new data streams against reliable metrics. True success will depend on the use of novel techniques to measure positive changes in population health.[42]

COMPETING INTEREST STATEMENT

MD and MJP hold equity in Sickweather Inc. MD has received consulting fees from Bloomberg LP, and holds equity in Good Analytics Inc. These organizations did not have any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. All other authors declare no competing interests.

FUNDING STATEMENT

Preparation of this manuscript was supported in part by the National Institute of General Medical Sciences under award number R01GM114771 to DAB and SCQ and by the National Science Foundation under award number IIS-1657338 to XH and MJP.

CONTRIBUTORSHIP STATEMENT

XH, MCS, DAB, MD, SCQ, and MJP contributed to the design of the study. XH, JC, MD, and MJP contributed to data collection. XH, MCS, JC, DAB, and MJP performed data analysis. XH, AMJ, DAB, SCQ, and MJP interpreted the results. All authors contributed to the editing of this manuscript.

DATA SHARING STATEMENT

All Twitter data used in this study is available in the following repository: https://figshare.com/account/projects/31742/articles/6213878

This contains the annotations for training the classifiers, as well as the classifier inferences on the full dataset. This also contains the extracted metadata, including demographics and location. In accordance with the Twitter terms of service, raw tweets are not shared, but identifiers are shared which can be used to download the tweets.

ACKNOWLEDGEMENT

An early version of this research was presented at the AAAI Joint Workshop on Health Intelligence (W3PHIAI) in February 2017.

REFERENCES

[1] L. A. Grohskopf *et al.*, “Prevention and Control of Seasonal Influenza With Vaccines: Recommendations of the Advisory Committee on Immunization Practices—United States, 2017–18 Influenza Season,” *Am. J. Transplant.*, vol. 17, no. 11, pp. 2970–2982, 2017.

[2] CDC, “Morbidity and Mortality Weekly Report (MMWR),” 2017. [Online]. Available: https://www.cdc.gov/mmwr/volumes/66/rr/rr6602a1.htm. [Accessed: 08-Mar-2018].

[3] T. Santibanez *et al.*, “Flu Vaccination Coverage, United States, 2016-17 Influenza Season,” 2017. [Online]. Available: https://www.cdc.gov/flu/fluvaxview/coverage-1617estimates.htm. [Accessed: 09-Mar-2018].

[4] CDC, “Influenza Vaccination Coverage | FluVaxView | Seasonal Influenza | CDC,” *Centers for Disease Control and Prevention*, Dec-2017. [Online]. Available: https://www.cdc.gov/flu/fluvaxview/index.htm. [Accessed: 09-Mar-2018].

[5] S. Keeter, “The Impact of Cell Phone Noncoverage Bias on Polling in the 2004 Presidential Election,” *Public Opin. Q.*, vol. 70, no. 1, pp. 88–98, Jan. 2006.

[6] R. Iachan, C. Pierannunzi, K. Healey, K. J. Greenlund, and M. Town, “National weighting of data from the Behavioral Risk Factor Surveillance System (BRFSS),” *BMC Med. Res. Methodol.*, vol. 16, no. 1, p. 155, Nov. 2016.

[7] N. V. P. Office, “Flu Vaccination Trends,” *US Department of Health and Human Services*, 2017. [Online]. Available: https://www.hhs.gov/nvpo/resources/flu/index.html.

[8] D. A. Broniatowski, M. J. Paul, and M. Dredze, “National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic,” *PLoS One*, vol. 8, no. 12, 2013.

[9] E. VELASCO, T. AGHENEZA, K. DENECKE, G. KIRCHNER, and T. I. M. ECKMANNS, “Social Media and Internet-Based Data in Global Systems for Public Health Surveillance: A Systematic Review,” *Milbank Q.*, vol. 92, no. 1, pp. 7–33, 2014.

[10] L. E. Charles-Smith *et al.*, “Using social media for actionable disease surveillance and outbreak management: a systematic literature review,” *PLoS One*, vol. 10, no. 10, p. e0139701, 2015.

[11] C. Corley, D. Cook, A. Mikler, and K. Singh, “Text and Structural Data Mining of Influenza Mentions in Web and Social Media,” *Int. J. Environ. Res. Public Health*, vol. 7, no. 12, pp. 596–615, Feb. 2010.

[12] N. Collier, N. Son, and N. Nguyen, “OMG U got flu? Analysis of shared health messages for bio-surveillance,” *J. Biomed. Semantics*, vol. 2, no. Suppl 5, p. S9, 2011.

[13] A. Odone *et al.*, “Effectiveness of interventions that apply new media to improve vaccine uptake and vaccine coverage,” *Hum. Vaccin. Immunother.*, vol. 11, no. 1, pp. 72–82, 2015.

[14] M. Dredze, D. A. Broniatowski, and K. M. Hilyard, “Zika vaccine misconceptions: A social media analysis,” *Vaccine*, vol. 34, no. 30, pp. 3441–3442, Jun. 2016.

[15] G. A. Powell *et al.*, “Media content about vaccines in the United States and Canada, 2012–2014: An analysis using data from the Vaccine Sentimeter,” *Vaccine*, vol. 34, no. 50, pp. 6229–6235, 2016.

[16] G. J. Kang *et al.*, “Semantic network analysis of vaccine sentiment in online social media,” *Vaccine*, vol. 35, no. 29, pp. 3621–3638, 2017.

[17] M. Salathé and S. Khandelwal, “Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control,” *PLOS Comput. Biol.*, vol. 7, no. 10, p. e1002199, Oct. 2011.

[18] M. Salathé, D. Q. Vu, S. Khandelwal, and D. R. Hunter, “The dynamics of health behavior sentiments on a large online social network,” *EPJ Data Sci.*, vol. 2, no. 1, p. 4, 2013.

[19] E. J. Nelson, J. Hughes, J. M. Oakes, J. S. Pankow, and S. L. Kulasingam, “Estimation of Geographic Variation in Human Papillomavirus Vaccine Uptake in Men and Women: An Online Survey Using Facebook Recruitment,” *J. Med. Internet Res.*, vol. 16, no. 9, p. e198, Sep. 2014.

[20] A. G. Dunn, D. Surian, J. Leask, A. Dey, K. D. Mandl, and E. Coiera, “Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States,” *Vaccine*, vol. 35, no. 23, pp. 3033–3040, May 2017.

[21] Z. Tufekci, “Big Questions for Social Media Big Data: Representativeness, Validity and Other Methodological Pitfalls.,” in *ICWSM*, 2014, vol. 14, pp. 505–514.

[22] R. Cohen and D. Ruths, “Classifying political orientation on Twitter: It’s not easy!,” in *ICWSM*, 2013.

[23] M. J. Paul and M. Dredze, “Discovering Health Topics in Social Media Using Topic Models,” *PLoS One*, vol. 9, no. 8, p. e103408, Aug. 2014.

[24] M. Dredze, M. J. Paul, S. Bergsma, and H. Tran, “Carmen: A twitter geolocation system with applications to public health,” in *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, 2013, vol. 23, p. 45.

[25] R. Knowles, J. Carroll, and M. Dredze, “Demographer: Extremely simple name demographics,” in *Proceedings of the First Workshop on NLP and Computational Social Science*, 2016, pp. 108–113.

[26] National Center for Immunization and Respiratory Diseases, “National Immunization Surveys (NIS),” 2018. .

[27] National Center for Chronic Disease Prevention and Health Promotion, “Behavioral Risk Factor Surveillance System Questionaires,” 2018. .

[28] C. Callison-Burch and M. Dredze, “Creating speech and language data with Amazon’s Mechanical Turk,” in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon’s Mechanical Turk*, 2010, pp. 1–12.

[29] X. Huang *et al.*, “Examining patterns of influenza vaccination in social media,” in *AAAI Joint Workshop on Health Intelligence (W3PHIAI)*, 2017, pp. 542–546.

[30] J. Franke, W. K. Härdle, and C. M. Hafner, “ARIMA Time Series Models,” in *Statistics of Financial Markets: An Introduction*, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 255–282.

[31] B. Efron and R. Tibshirani, “[Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy]: Rejoinder,” *Stat. Sci.*, vol. 1, no. 1, pp. 77–77, Feb. 1986.

[32] L. Flekova, J. Carpenter, S. Giorgi, L. Ungar, and D. Preo\ctiuc-Pietro, “Analyzing biases in human perception of user age and gender from text,” in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, vol. 1, pp. 843–854.

[33] J. Du, J. Xu, H.-Y. Song, and C. Tao, “Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data.,” *BMC Med. Inform. Decis. Mak.*, vol. 17, no. Suppl 2, p. 69, 2017.

[34] T. R. Tangherlini *et al.*, "'Mommy Blogs' and the Vaccination Exemption Narrative: Results From A Machine-Learning Approach for Story Aggregation on Parenting Social Media Sites.," *JMIR public Heal. Surveill.*, vol. 2, no. 2, p. e166, 2016.

[35] X. Zhou, E. W. Coiera, G. Tsafnat, D. Arachi, M.-S. Ong, and A. G. Dunn, "Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter," *Stud. Health Technol. Inform.*, vol. 216, pp. 761–765, 2015.

[36] K. A. McGregor and M. E. Whicker, "50 - Natural Language Processing Approaches to Understand HPV Vaccination Sentiment," *J. Adolesc. Heal.*, vol. 62, no. 2, Supplement, pp. S27–S28, 2018.

[37] K. P. Alberti, J. P. Guthmann, F. Fermon, K. D. Nargaye, and R. F. Grais, "Use of Lot Quality Assurance Sampling (LQAS) to estimate vaccination coverage helps guide future vaccination efforts," *Trans. R. Soc. Trop. Med. Hyg.*, vol. 102, no. 3, pp. 251–254, 2008.

[38] M. Wagner, V. Lampos, E. Yom-Tov, R. Pebody, and I. J. Cox, "Estimating the Population Impact of a New Pediatric Influenza Vaccination Program in England Using Social Media Content," *J. Med. Internet Res.*, vol. 19, no. 12, p. e416, Dec. 2017.

[39] J. M. Krogstad, "Social media preferences vary by race and ethnicity," *Pew Research Center*, Feb-2015. [Online]. Available: http://www.pewresearch.org/fact-tank/2015/02/03/social-media-preferences-vary-by-race-and-ethnicity/. [Accessed: 08-Mar-2018].

[40] CDC, "Flu Vaccination Coverage, United States, 2016-17 Influenza Season," 2017. [Online]. Available: https://www.cdc.gov/flu/fluvaxview/coverage-1617estimates.htm#age-group-adults. [Accessed: 08-Mar-2018].

[41] HealthyPeople, "Immunization and Infectious Diseases," *HealthyPeople.gov*. [Online]. Available: https://www.healthypeople.gov/2020/topics-objectives/topic/immunization-and-infectious-diseases. [Accessed: 09-Mar-2018].

[42] S. J. Mooney, D. J. Westreich, and A. M. El-Sayed, "Epidemiology in the Era of Big Data," *Epidemiology*, vol. 26, no. 3, pp. 390–394, May 2015.