ORIGINAL RESEARCH & CONTRIBUTIONS

Sentiment, Contents, and Retweets: A Study of Two Vaccine-Related Twitter Datasets

Elizabeth B Blankenship, MPH; Mary Elizabeth Goff, MPH; Jingjing Yin, PhD; Zion Tsz Ho Tse, PhD; King-Wa Fu, PhD; Hai Liang, PhD; Nitin Saroha, MS; Isaac Chun-Hai Fung, PhD

Perm J 2018;22:17-138

E-pub: 06/11/2018

https://doi.org/10.7812/TPP/17-138

ABSTRACT

Introduction: Social media platforms are important channels through which health education about the utility and safety of vaccination is conducted.

Objective: To investigate if tweets with different sentiments toward vaccination and different contents attract different levels of Twitter users' engagement (retweets).

Methods: A stratified random sample (N = 1425) of 142,891 #vaccine tweets (February 4, 2010, to November 10, 2016) was manually coded. All 201 tweets with 100 or more retweets from 194,259 #vaccineswork tweets (January 1, 2014, to April 30, 2015) were manually coded. Regression models were applied to identify factors associated with retweet frequency.

Results: Among #vaccine tweets, provaccine tweets (adjusted prevalence ratio = 1.5836, 95% confidence interval = 1.2130-2.0713, $p < 0.001$) and antivaccine tweets (adjusted prevalence ratio = 4.1280, 95% confidence interval = 3.1183-5.4901, $p < 0.001$) had more retweets than neutral tweets. No significant differences occurred in retweet frequency for content categories among antivaccine tweets. Among 411 links in provaccine tweets, Twitter (53; 12.9%), content curator Trap.it (14; 3.4%), and the Centers for Disease Control and Prevention (8; 1.9%) ranked as the top 3 domains. Among 325 links in antivaccine tweets, social media links were common: Twitter (44; 14.9%), YouTube (25; 8.4%), and Facebook (10; 3.4%). Among highly retweeted #vaccineswork tweets, the most common theme was childhood vaccinations (40%; 81/201); 21% mentioned global vaccination improvement/efforts (42/201); 29% mentioned vaccines can prevent outbreaks and deaths (58/201).

Conclusion: Engaging social media key opinion leaders to facilitate health education about vaccination in their tweets may allow reaching a wider audience online.

INTRODUCTION

Communicating the benefits of vaccination to the public remains a challenge amid the presence of the antivaccination movement.[1] This movement causes hesitance and criticism among parents regarding vaccines for myriad reasons, including lack of trust in government and the pharmaceutical industry, feared acute and long-term side effects, and concern over the chemical makeup of the vaccines themselves.[2,3] Outbreaks of vaccine-preventable diseases in the US occur more often as rates of vaccination decline. For example, measles had been eliminated in the US since 2000 until travel-related imported cases led to outbreaks in recent years, including a large outbreak among unvaccinated Amish individuals in 2014.[4]

Social media has become a major mode of global communication, through which dissemination of information is easier than ever. Currently, 21% of all US adults use Twitter, with 42% of those users visiting the Twitter platform daily.[5] With more than 328 million users,[6] Twitter is a convenient tool for discussing public health topics, including vaccination. Both provaccine and antivaccine information is prevalent on Twitter. Understanding how vaccine-related information disseminates on Twitter is vital, especially because a minority of users are openly skeptical about vaccines and advocate against vaccination. Prior research focused on how vaccines were portrayed on social media[7,8] and how misinformation or controversial information spread.[9-11] Researchers attempted to develop methods to monitor vaccination sentiment in real time by primarily focusing on the incidence of tweets with positive and negative sentiments over time.[5,12] Efforts were made to use supervised machine learning methods to predict a tweet's sentiment toward vaccination, using either contents of manually coded tweets or their users' connections as classifiers.[13,14] Although important progress has been made, questions remain at the microlevel, such as whether tweets containing provaccine or antivaccine sentiment and information attract attention on Twitter.

Here, we provide definitions to a few Twitter-specific terms. *Retweets* are tweets that users repost after reading them in their timeline.[15] A Twitter user's *follower count* is the number of Twitter users who follow the account of a user. A Twitter user's *friend count* is the number of Twitter users whom the user follows on Twitter. A Twitter user's *status count* is the number of status updates (tweets) that the user has posted so far. A Twitter user's *favorite count* is the number of likes the user has ever given to other people's tweets.

In this article, we report analyses of two distinct datasets that, in turn, addressed four interrelated research questions.

Elizabeth B Blankenship, MPH, was a Graduate Student in Epidemiology and Environmental Health Sciences at the Jiann-Ping Hsu College of Public Health at Georgia Southern University in Statesboro. E-mail: eb03763@georgiasouthern.edu. **Mary Elizabeth Goff, MPH,** was a Graduate Student in Epidemiology and Environmental Health Sciences at the Jiann-Ping Hsu College of Public Health at Georgia Southern University in Statesboro. E-mail: eg01587@georgiasouthern.edu. **Jingjing Yin, PhD,** is an Assistant Professor of Biostatistics at the Jiann-Ping Hsu College of Public Health at Georgia Southern University in Statesboro. E-mail: jyin@georgiasouthern.edu. **Zion Tsz Ho Tse, PhD,** is an Associate Professor in the School of Electrical and Computer Engineering at the College of Engineering at the University of Georgia in Athens. E-mail: ziontse@uga.edu. **King-Wa Fu, PhD,** is an Associate Professor at the Journalism and Media Studies Centre at the University of Hong Kong and a Visiting Associate Professor at the Massachusetts Institute of Technology Media Lab in Cambridge. E-mail: kwfu@hku.hk. **Hai Liang, PhD,** is an Assistant Professor in the School of Journalism and Communication at the Chinese University of Hong Kong. E-mail: hailiang@cuhk.edu.hk. **Nitin Saroha, MS,** is a Graduate Student in Computer Science at the University of Georgia in Athens. E-mail: ns10510@uga.edu. **Isaac Chun-Hai Fung, PhD,** is an Assistant Professor in Epidemiology and Environmental Health Sciences at the Jiann-Ping Hsu College of Public Health at Georgia Southern University in Statesboro. E-mail: cfung@georgiasouthern.edu.

Study A: #vaccine Twitter Dataset

In Study A, we analyzed a 1% stratified random sample of a corpus of tweets with the hashtag #vaccine, a hashtag used by both provaccine and antivaccine advocates. We believed that tweets carrying stronger sentiments would attract more attention and retweets from those who wanted to share them. Therefore, we hypothesized as follows:

> Hypothesis 1: Antivaccine and provaccine #vaccine tweets differ in their retweet count, compared with tweets of neutral sentiment.

We also postulated that users' characteristics could be potential confounders in the association between sentiment and retweet frequency, and therefore we included the users' follower count, friend count, status count, and favorite count in our analysis.

We also speculated whether different categories of antivaccine contents attracted different quantities of retweets.

> Hypothesis 2: Different categories of contents among antivaccine #vaccine tweets differ in their retweet count.

We were also interested in the source of information in the provaccine and antivaccine #vaccine tweets.

Research Question 1: What were the embedded Uniform Resource Locator (URL) domains in the provaccine and antivaccine #vaccine tweets?

Study B: #vaccineswork Twitter Dataset

In Study B, we analyzed a corpus of tweets with the hashtag #vaccineswork. This hashtag has been used by public health professionals when they promoted vaccination.[16] Because the distribution of retweet count is highly skewed with only very few tweets having high retweet count, it is likely that tweets with high retweet counts are read by many and may have influence over the knowledge, attitudes, or perceptions of many users, whereas tweets with few retweets do not. Given the need to perform manual coding, in Study B, we chose to focus our limited resources on tweets that carry the greatest influence rather than tweets with minimal influence. We manually categorized the contents of tweets containing #vaccineswork that were retweeted 100 or more times. We provided a descriptive analysis of the distribution of topics among this sample of highly retweeted tweets. We also combined several topics into a categorical variable and tested if statistical association existed between retweet frequency and that categorical content variable.

Research Question 2: Would highly retweeted provaccine contents on Twitter (#vaccineswork tweets) differ by content in their retweet frequency?

METHODS

This research was approved by the institutional review board of Georgia Southern University (H15083) under the B2 exempt category because the social media posts analyzed in this study are considered publicly observable behavior.

Study A: #vaccine Twitter Dataset

Data

The #vaccine dataset was retrieved using Twitter Application Programming Interface (API; Online Supplementary Materials[a]). The data contain 142,891 tweets from Twitter with the hashtag #vaccine, from February 4, 2010, to November 10, 2016 (inclusive). Retweet frequency and other meta-data reported in this paper were correct as of the data retrieval date (November 10, 2016). Data were then stratified by month, and a random 1% sample of tweets was collected from each month, resulting in the extraction of 1425 tweets for manual coding.

Manual Coding

Authors MEG and EBB previewed tweets for recurring themes within the content of the tweets and developed a codebook (with example tweets) on the basis of these themes. The codebook is available in the Online Supplementary Materials.[a] Following the codebook, MEG and EBB independently, manually coded the contents of the tweets. Each content category was manually coded as a binary variable (0 = no, 1 = yes). Tweets were coded into the following sentiment categories: Provaccine sentiment, neutral sentiment, and antivaccine sentiment. Provaccine sentiment refers to tweets that explicitly communicated to readers that a vaccine is a safe and effective way of preventing diseases. Antivaccine sentiment refers to tweets that expressed skepticism or denial of vaccines as a safe and effective way of preventing diseases. Neutral sentiment refers to tweets with plain statements related to vaccine, such as its availability. Sentiment categories were merged into one categorical variable (1 = Neutral, 2 = Positive, 3 = Negative). Tweets that were deemed irrelevant or whose sentiment could not be determined (n = 81) were removed from further analysis. A total of 1344 tweets in English with categorized sentiments were analyzed (1326 were labeled as English in the Twitter metadata; 18 were labeled otherwise but were found to be in English through manual coding). Tweets that were identified as "antivaccine" (n = 325) were further manually coded into 2 themes that are not mutually exclusive (each being a binary variable): 1) perceived harmful risks, alleged side effects and/or deaths caused by vaccines (eg, autism, seizures, fatalities); and 2) distrust of government, pharmaceutical companies, scientists, and organizations that support vaccination efforts (eg, the Bill & Melinda Gates Foundation). Any antivaccine tweets that did not fall into either of the 2 themes were labeled as miscellaneous (tweets that are antivaccine but do not meet any of the content categories). Examples are given in the codebook in the Online Supplementary Materials.[a]

Statistical Analysis and Resolving URL

All statistical analyses in this experiment were performed in R Version 3.2.2 or 3.3.0 (R Foundation, Vienna, Austria). Negative binomial regression models were used because of overdispersion of the retweet frequency in this dataset. Because we postulated that the users' characteristics could potentially be confounders to the statistical association between sentiment toward vaccine and retweet frequency, the users' followers count, friends count, status count, and favorite count were included in our analysis. Given the highly skewed distributions of these variables, we converted these continuous variables into binary variables for better interpretation. The data were dichotomized as either below the geometric mean (labeled as 0) or not (labeled as 1). The cutoff value of $\alpha = 0.05$ was chosen a priori for statistical significance. The short URLs of provaccine sentiment tweets and antivaccine sentiment tweets were resolved using R to their original URLs, and we extracted their domains. Descriptive statistics for URL domains that appeared 3 times or more are presented in the article.

Study B: #vaccineswork Twitter Dataset

Data

The data used for this study were purchased through GNIP Inc, which is a subsidiary of Twitter Inc in Boulder, CO. The dataset contained all tweets with the hashtag #vaccineswork from January 1, 2014, to April 30, 2015. The original dataset contained 194,259 tweets. Tweets therein with a threshold of greater than or equal to 100 retweets were grouped by subset from the original dataset for further analysis (N = 201). Retweet frequency and other metadata reported in this article were correct as of the date of data retrieval from GNIP Inc (early May 2015).

Manual Coding

Authors EBB and MEG developed a codebook by previewing the data for recurring themes. The codebook contained the following content categories: Mention of deaths and/or outbreaks of diseases that are vaccine-preventable; child vaccinations; mention of professional organizations, such as the Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO); mention of vaccine efficacy; mention of global vaccination importance; mention of people lacking access to vaccinations; tweets referring to World Immunization Awareness Week; mention of outbreaks of vaccine-preventable diseases; and provaccine statements directed at antivaccination sentiment. Tweets that did not meet any of these content categories were coded as miscellaneous. Content categories were not mutually exclusive (ie, the content of a tweet can be coded as "yes" in more than 1 category). Each content category was coded as a binary variable (0 = no, 1 = yes). Both EBB and MEG independently, manually coded all 201 tweets. Interrater reliability between the 2 coders was assessed by analyzing Cohen κ for each content category. The κ values for all content categories were > 0.8, implying a good interrater reliability.

The corresponding author (ICHF) further combined the content categories of "Mentions vaccines preventing deaths and/or outbreaks" and/or "Mentions efficacy of vaccines" into one category (Category 1), and those of "Mentions child vaccination" and/or "Mentions global vaccination improvement/efforts" into a single category (Category 2). Any tweet that was coded "yes" for Categories 1 and 2 was coded as Category 3, and any tweet that did not fall into Category 1 or 2 was coded as Category 0. A new categorical variable of content was thus created (see Online Supplementary Materials[a]).

Statistical Analysis

All statistical analysis was performed using R version 3.2.2 or 3.3.0. Retweet frequency in this dataset of 201 manually coded tweets was overdispersed and truncated with a theoretical minimum value of 100. Therefore, a zero-truncated, negative binomial regression model was applied to new outcome variables[17]: Retweet truncated = Retweet frequency - 99. The regression model was applied after removing 4 apparent outliers from our dataset (bringing the total to 197 tweets). The cutoff value of $\alpha = 0.05$ was chosen a priori for statistical significance.

RESULTS

Study A: #vaccine

Of the sample of 1344 #vaccine tweets that were coded with vaccine-related sentiments, provaccine tweets accounted for 32.4% (436/1344) of the sample, neutral tweets for 43.4% (583/1344), and antivaccine tweets for 24.2% (325/1344; Table 1). Regarding the proportion of tweets with URL links therein, there was no statistically significant difference ($\chi^2 = 4.4297$, degrees of freedom = 2, p = 0.1092). In the

Table 1. Frequency of occurrence of each binary content category variable of the 1% stratified random sample of #vaccine population of tweets (n = 1344)[a]

Content category of tweet	Frequency (%)	Frequency (%) of tweets with URL in the category	Example
Descriptive statistics of the 1% stratified random sample of #vaccine tweets			
All	1344 (100)	621 (46.2)	—
Provaccine sentiment	436 (32.4)	191 (43.8)	Gardasil the vaccine that can prevent #cervical cancer in girls won FDA blessing as #vaccine to #prevent #anal #cancer
Neutral sentiment	583 (43.4)	238 (40.8)	India set to release its first #H1N1 #SwineFlu #vaccine
Antivaccine sentiment	325 (24.2)	192 (59.1)	Another Childhood #Vaccine Link to #Autism: Dr. Andrew Wakefield is proven right again
Descriptive statistics of the sample of 325 antivaccine tweets			
All	325 (100)	192 (59.1)	—
Mentions perceived risks and/or dangers of vaccines	153 (47.1)	133 (86.9)	Check this video out-Doctor Admits #Vaccine Is More #Deadly Than #Swine #Flu & wont give 2 his #children. http://t.co/aPrVEmR via @youtube
Distrust of government, pharmaceutical companies, scientists, etc.	85 (26.2)	70 (82.4)	@MedPedsDoctor Why did the #FDA give #vaccine manufacturers blanket immunity from any and all defects in manufactur [sic] then?
Both categories	54 (16.6)	44 (81.5)	#GSK Fined for Killing 14 Babies in #Vaccine Trial - YouTube http://t.co/0XdezKTl
Not in either category ("miscellaneous")	33 (10.2)	25 (75.8)	#Flu #Vaccines Don't Work 99% of the Time http://bit.ly/9vXtZk #vaccine #junkscience #bigpharma

[a] Tweets in provaccine categories and neutral categories are mutually exclusive. Subcategories of antivaccine sentiment are not mutually exclusive. An antivaccine tweet may meet the criteria of multiple subcategories. Tweets deemed irrelevant or whose sentiment was unable to be determined (n = 81) were removed from further analysis. Thus, the total number of #vaccine tweets included in this analysis was 1344.

FDA = Food and Drug Administration; URL = Uniform Resource Locator.

antivaccine subcorpus of tweets (n = 325), 153 (47.1%) tweets mentioned only perceived risks and/or dangers of vaccines; 85 (26.2%) tweets mentioned only distrust of scientific entities such as the government, pharmaceutical companies, and scientists; 54 (16.6%) tweets mentioned both themes; and 33 (10.2%) tweets did not fit into either of the 2 themes ("miscellaneous"; Table 1). No significant differences in the proportion of tweets with URL links therein were observed among the 4 categories (χ^2 = 3.0012, degrees of freedom = 3, p = 0.3914).

In Table 2, we present the descriptive statistics of the retweet frequency, and the counts of users' followers, friends, status updates, and favorites. These data were very skewed. For example, the median for retweet frequency for the sample and those for subsamples for positive, neutral, and negative sentiments were 0. For the users' characteristics, the means were much larger than the medians. Therefore, for subsequent analysis, we dichotomized the users' characteristics data as below the geometric mean or not, and thus converted the continuous variables into binary variables.

First, in the univariate analysis, both provaccine and antivaccine tweets had statistically significantly more retweets than neutral tweets; the users' follower count, friend count, and status count were found to have statistically significant associations with retweet frequency (Table 3). In the multivariable regression analysis, provaccine tweets had 1.58 times as many retweets as neutral tweets (adjusted prevalence ratio = 1.5836, 95% confidence interval [CI] = 1.2130-2.0713, p < 0.001), and antivaccine tweets had 4.13 times as many retweets as neutral tweets (adjusted prevalence ratio = 4.1280, 95% CI = 3.1183-5.4901, p < 0.001) after controlling for users' follower count, friend count, and status count (Table 3). Thus, antivaccine and provaccine #vaccine tweets differed in their retweet count, compared with tweets of neutral sentiment. Antivaccine tweets received more retweets than did provaccine tweets and neutral tweets. It is important to note that the retweet frequency of tweets posted by users with high follower count was 3.88 times (adjusted prevalence ratio = 3.8771; 95% CI = 2.9977-5.0295; p < 0.001) that of users with low follower count. To the contrary, users with high status count (ie, number of tweets ever tweeted) had 24% fewer retweets (prevalence ratio = 0.7597; 95% CI = 0.5856-0.9824; p = 0.033) than did users with low count of status updates.

Second, among the antivaccine subcorpus of tweets (n = 325), univariate negative binomial regression found that there were no significant differences between tweets that mentioned perceived risks and/or dangers of vaccines and those that did not (prevalence ratio = 0.74,

Table 2. Descriptive statistics of the 1% stratified random sample of the #vaccine population of tweets (n = 1344)

Parameter	Mean	Median	25%	75%	Minimum	Maximum
Retweet count						
All	1.15997	0	0	1	0	60
Provaccine	0.96560	0	0	1	0	26
Neutral	0.60720	0	0	1	0	19
Antivaccine	2.41231	0	0	2	0	60
Users' follower count						
All	9666.50	1451.5	425.5	4368.75	9	1,028,118
Provaccine	9339.97	1351	409.75	4491	9	1,028,118
Neutral	7067.41	1345	361	3102.5	9	415,830
Antivaccine	14,766.92	2105	689	6182	11	247,004
Users' friend count						
All	3651.98	771	302.75	2054.75	0	157,038
Provaccine	2075.51	696.5	280	1836	0	121,919
Neutral	2306.17	698	231.5	1664	0	154,537
Antivaccine	8181.03	1362	463	3976	0	157,038
Users' status count						
All	32,761.01	8763	3397.25	28,982.5	9	1,030,714
Provaccine	19,748.03	7355.5	2742	15,799.5	26	871,221
Neutral	28,559.73	8208	2718	23,144	9	821,676
Antivaccine	57,754.87	22,638	6521	50,022	101	1,030,714
Users' favorite count						
All	3936.63	263.5	12.75	1433	0	283,691
Provaccine	1946.40	269.5	29.75	1210.75	0	59,735
Neutral	2236.35	113	1	882	0	283,691
Antivaccine	9656.62	873	69	6067	0	104,597

Table 3. Univariate analysis showing prevalence ratio of retweet frequency of provaccine and antivaccine tweets relative to neutral tweets in #vaccine sample after excluding tweets that were not relevant or whose sentiment could not be determined (n = 1344)

	Univariate model		Multivariable model	
Predictor	Prevalence ratio (95% CI)	p value	Prevalence ratio (95% CI)	p value
Sentiment of tweets (categorical variable)				
Neutral	Reference	—	Reference	—
Provaccine	1.5902 (1.2075-2.0988)	0.001	1.5836 (1.2130-2.0713)	< 0.001
Antivaccine	3.9728 (2.9854-5.3164)	< 0.001	4.1280 (3.1183-5.4901)	< 0.001
Users' characteristics (binary variables)				
Follower count	3.4619 (2.7392-4.3794)	< 0.001	3.8771 (2.9977-5.0295)	< 0.001
Friend count	1.3295 (1.0400-1.6980)	0.023	0.8946 (0.7008-1.1389)	0.402
Status count	1.7140 (1.3456-2.1844)	< 0.001	0.7597 (0.5856-0.9824)	0.033
Favorite count	1.1895 (0.9280-1.5218)	0.169	Not included in the model	—

CI = confidence interval.

95% CI = 0.47-1.15, p = 0.20), and between tweets that mentioned distrust of government, pharmaceutical companies, scientists, and so on, and those that did not (prevalence ratio = 1.00, 95% CI = 0.65-1.56, p = 0.99). Thus, our hypothesis that different categories of contents among antivaccine #vaccine tweets differ in their retweet count was rejected.

Third, a total of 411 URL links were identified in 436 provaccine tweets: 36 tweets had 2 URLs, and 339 tweets had 1 URL. Among these links, Twitter (53; 12.9%), content curator Trap.it (14; 3.4%), and the CDC (8; 1.9%) were the top 3 domains. A total of 296 URL links were identified in 325 antivaccine tweets: 24 tweets had 2 URLs, and 248 had 1. Among these links, 26.7% of them were links to other tweets (44; 14.9%), YouTube videos (25; 8.4%), or Facebook (10; 3.4%). There were long tails with low frequency (1 or 2) for the URL domain frequency distributions among both provaccine and antivaccine tweets. Tables S1 and S2 in the Online Supplementary Materials[a] detail the URL domains of URL links identified among provaccine and antivaccine #vaccine tweets.

Study B: #vaccineswork Twitter Dataset

Among our sample of 201 #vaccineswork tweets with 100 retweets or more, the most common theme observed was childhood vaccinations (40%; 81/201; Table 4). One in 5 tweets mentioned the global vaccination improvement/efforts (21%; 42/201). Nearly 3 in 10 tweets mentioned how vaccines can prevent outbreaks and deaths (29%; 58/201), 18% (37/201) mentioned a professional organization (eg, WHO or CDC), and 18% (36/201) discussed the efficacy of vaccines and vaccination of the population. Fifteen percent (31/201) mentioned a certain group of people (ie, a population, race/ethnicity, and/or country) and their lack of access to vaccines and routine vaccination; 12% (24/201) of tweets mentioned outbreaks and/or deaths that were caused by vaccine-preventable diseases; 10% (20/201) of tweets were focused on World Immunization Awareness Week; and 6% (13/201) of tweets were provaccination stances directed toward antivaccination sentiment (Table 4).

As previously described, some categories were dropped and others merged to create a categorical variable of 2 mutually exclusive categories and their combination for further regression analysis. After removing 4 outliers, the univariable zero-truncated negative binomial regression model was applied to the dataset (n = 197). No statistically significant association was observed between the categorical variable of combined categories and retweet frequency (Table S3 in Online Supplementary Materials[a]).

Here, we described the 4 outliers that were most retweeted in our #vaccineswork dataset. The most retweeted tweet in the dataset was tweeted by American politician Hillary Clinton: "The science is clear: The earth is round, the sky is blue, and #vaccineswork. Let's protect all our kids. #GrandmothersKnowBest." The tweet was retweeted 33,164 times at the time when the dataset was purchased.

The second most retweeted tweet in this dataset was "*drops microphone* #antivax #vaccineswork #VaccinateYourKids http://t.co/1Nysbfkh7N" (retweet frequency = 5032). It was tweeted by @DocBastard, who described himself as a trauma surgeon in his user profile. This tweet ended with a link to an image of another physician's social media post about how he handled parents who declined to have their children vaccinated on schedule.

The third most retweeted tweet was tweeted by the WHO (@WHO): "World Immunization Week starts today! Close the immunization gap, #VaccinesWork http://t.co/G3CjZKdhyv http://t.co/kZkiYPdEan" (retweet frequency = 1368). The first link in the tweet takes the user to a page on the WHO Web site about World Immunization Week. The second link takes the user to an infographic by the WHO that states, "Today 1 in 5 children worldwide is missing out on vital immunization."

The fourth most retweeted tweet was tweeted by Sue Desmond-Hellmann, MD, MPH, the Chief Executive Officer of the Bill & Melinda Gates Foundation: "It's impossible to argue with results like this. #vaccineswork http://t.co/yOeDVi2m0s" (User: @Sue Desmond-Hellmann; retweet frequency = 1182). The link therein takes the user to the tweet with an infographic that describes the decrease in percentage of annual morbidity of vaccine-preventable diseases in the US from the prevaccine era to the present.

Table 4. Frequency of occurrence of each binary content category variable for the sample of #vaccineswork tweets with 100 or more retweets (N = 201)

Content category	Frequency (%)
Mentions vaccines preventing deaths and/or outbreaks	58 (28.86)
Mentions child vaccination	81 (40.30)
Mentions a professional organization	37 (18.41)
Mentions efficacy of vaccines	36 (17.91)
Mentions global vaccination improvement/efforts	42 (20.90)
Mentions people's lack of access to vaccines and vaccination	31 (15.42)
Focuses on World Immunization Awareness Week	20 (9.95)
Mentions outbreaks and/or deaths of vaccine-preventable diseases	24 (11.94)
Provaccine statements aimed toward antivaccination sentiment	13 (6.47)

DISCUSSION

In this study, we analyzed two datasets of vaccine-related tweets. We investigated the retweet frequency of a random sample of tweets within the #vaccine corpus, as well as the retweet frequency of a sample of highly retweeted tweets in the #vaccineswork corpus.

Among our random sample of #vaccine tweets, antivaccine tweets were retweeted more often, receiving 4.13 times as many retweets as neutral tweets, whereas provaccine tweets received 1.58 times as many retweets as neutral tweets. No differences in retweet frequency were observed for tweets carrying 2 content categories of antivaccine contents and those that did not.

Childhood vaccination appeared to be one of the most frequent topics in the #vaccineswork sample, with approximately 40% of the dataset mentioning childhood vaccination. This could be because of the increased interest in childhood vaccinations (eg, the number of vaccinations

necessary, whether they are necessary at all, or their importance) in 2014 to 2015.[18] Other top conversations in this corpus discussed the improvement in global vaccination and how vaccines can prevent outbreaks or deaths owing to vaccine-preventable diseases.

One of our key findings is that despite the provaccine health communication efforts made by public health agencies, as far as #vaccine tweets are concerned, on a tweet-by-tweet basis, antivaccine tweets may be receiving more attention (as reflected in the number of retweets) than provaccine tweets or neutral tweets. A potential explanation is that although the supporters of the antivaccine movements are a minority in the population, many of them are very committed to their cause and are active online.[1] They retweeted tweets posted by like-minded individuals, forming an echo chamber.[11] A study by Bahk et al[12] found that antivaccine tweets persisted longer in a Twitter conversation about human papillomavirus than did the provaccine tweets. Our results added more evidence to the growing literature about the characteristics of antivaccine tweets.

The sources of information (URL domains) identified in the sample of #vaccine tweets can not only help public health professionals understand through which platforms people are gathering their information about vaccines but also can provide insight to what platforms or sites professionals should target when disseminating provaccine information. Given the use of Twitter across the opinion spectrum, it is not surprising that the top URL domain for both provaccine and antivaccine tweets was Twitter itself. In fact, it might reflect the growing trends that individuals rely on social media as their main source of news and information, compared with direct visits to Web sites of media or health organizations.[19] Regarding URL domains in provaccine tweets in this corpus, many were major news sources (eg, The Washington Post), public health agencies (cdc.gov), and Web sites that communicate science and medicine; some were from social media such as Facebook and Instagram. To the contrary, URL domains in antivaccine tweets included sources from social media sites (eg, Facebook and YouTube) as well as Internet news sources and Web sites that are skeptical of vaccines and the medical establishment, and that advocate individuals' right to decline vaccines for themselves and their children. Our results are congruent with the observed echo chamber effects on social media networks, in which people with similar ideas communicate with each other but not with people who disagree with them. As Del Vicario et al[11] showed with Facebook data, users consuming scientific news and conspiracy theories are usually two distinct polarized communities that are homogeneous among themselves. Bessi and colleagues[20] found that the debunking of conspiracy theories on Facebook were primarily read by users who frequently visited Facebook pages that shared scientific views and not by Facebook users who frequently consumed conspiracy theory Facebook posts; such observations cast doubt on the effectiveness of debunking conspiracy theories. A semantic network analysis of Internet articles shared by American Twitters users[21] found that Internet contents of antivaccine sentiment put great emphasis on children and institutions, including the CDC, the pharmaceutical industry, the medical profession, the mainstream media, and the state. Distrust of the industry and government agencies that communicate provaccine scientific messages was found to be the key underlying theme of the antivaccine Internet articles. Our results added further evidence to the literature that people with antivaccine sentiment obtain and share information from alternative sources, probably because of their distrust of public health, medical, and pharmaceutical establishments. Therefore, simply releasing more scientific information online through Web sites and social media may not help.[11]

The outliers of the #vaccineswork dataset suggested that having key opinion leaders who are active on social media to communicate our scientific message that "vaccines work" is important, as it is through them that provaccine messages can reach users who normally would not follow social media accounts of health agencies.

This study has some limitations. First, our samples were small. Given the labor-intensive nature of manual coding, we could not manually code every tweet in our corpora. In Study A, we analyzed a 1% stratified random sample of #vaccine tweets that was representative of the corpus. In Study B, we analyzed a sample of #vaccineswork tweets that were retweeted 100 or more times. Our analysis was meaningful because we covered the most retweeted, and thus the most influential, tweets.

Second, our original coding scheme in Study B provided useful insights, but the nonmutually exclusive categories rendered regression analysis difficult to interpret. Further analysis after dropping outliers from the dataset, and after dropping some themes and merging the others, found no statistical association between retweet frequency and combined themes. This can be potentially explained as a result of the study design, because we decided to focus on the most retweeted tweets and therefore could not identify any differences in retweet frequency between the combined themes. Future research may investigate other factors that might have an influence on retweet frequency of highly retweeted tweets, such as the temporal trends associated with the topic at the time (ie, a topic that is getting increased media coverage), and the topic that led to spikes in social media traffic (as in a case study of spikes of Chinese social media posts about 42 notifiable infectious diseases[22]).

Third, our analysis of URL links in the #vaccine sample in Study A was limited to their domains. For URL links to social media platforms such as Twitter and Facebook, we did not analyze the users who posted the original social media posts to which the tweet was linked, or the contents of such posts (which was the focus of recent studies such as in Kang et al[21]). Fourth, retweet frequency is only one of several metrics used to measure engagement of social media users with the original posts. Some fake accounts or Internet "bots" could artificially boost the retweet frequency of some tweets. We did not have access to information that would allow us to distinguish retweets by "bots" from retweets by genuine users. Fifth, our analyses were confined to two corpora of tweets with hashtags #vaccine

and #vaccineswork. Although this might limit the study's generalizability to other tweets, our analyses were able to focus on tweets that laid emphasis on vaccine (through the use of hashtags). Future research on tweets with and without other hashtags may enlighten us on the generalizability of our findings. Sixth, we retrieved our tweets with two English-language hashtags and, therefore, retrieved tweets that were predominantly in English. Future research can extend to investigate how Twitter users in different linguocultural communities responded to provaccine and antivaccine messages on Twitter. A recent study found that Twitter users who used different languages reacted differently to an outbreak.[23]

CONCLUSION

Among #vaccine tweets, antivaccine tweets attracted more engagement than did provaccine tweets. Antivaccine tweets and provaccine tweets were 4.1 and 1.6 times as likely, respectively, to be retweeted as were vaccine-related tweets with neutral sentiments. Among #vaccineswork tweets, we did not find evidence of differences in retweet frequency between themes. Reaching out to key opinion leaders on Twitter to promote provaccine messages may help reach Twitter users who would be otherwise unreached by public health agencies. ❖

[a] Online Supplementary Materials available at: www.thepermanentejournal.org/files/2018/17-138-Suppl.pdf

Disclosure Statement

Dr Fung received salary support from the Centers for Disease Control and Prevention (16IPA1609578). The data used in Study B were purchased using the start-up funds that Dr Fung received at the Jiann-Ping Hsu College of Public Health, Georgia Southern University, in Statesboro, GA. This article is not related to the Centers for Disease Control and Prevention-funded project of Dr Fung. The opinions expressed in this article do not represent the official positions of the Centers for Disease Control and Prevention or the US Government.

The author(s) have no conflicts of interest to disclose.

Acknowledgment

Kathleen Louden, ELS, of Louden Health Communications provided editorial assistance.

How to Cite this Article

Blankenship EB, Goff ME, Yin J, et al. Sentiment, contents, and retweets: A study of two vaccine-related twitter datasets. Perm J 2018;22:17-138. DOI: https://doi.org/10.7812/TPP/17-138

References

1. Kata A. Anti-vaccine activists, Web 2.0, and the postmodern paradigm—an overview of tactics and tropes used online by the anti-vaccination movement. Vaccine 2012 May 28;30(25):3778-89. DOI: https://doi.org/10.1016/j.vaccine.2011.11.112.
2. Luthy KE, Beckstrand RL, Callister LC, Cahoon S. Reasons parents exempt children from receiving immunizations. J Sch Nurs 2012 Apr;28(2):153-60. DOI: https://doi.org/10.1177/1059840511426578.
3. Dredze M, Broniatowski DA, Smith MC, Hilyard KM. Understanding vaccine refusal: Why we need social media now. Am J Prev Med 2016 Apr;50(4):550-2. DOI: https://doi.org/10.1016/j.amepre.2015.10.002.
4. Gastañaduy PA, Budd J, Fisher N, et al. A measles outbreak in an underimmunized Amish community in Ohio. N Engl J Med 2016 Oct 6;375(14):1343-54. DOI: https://doi.org/10.1056/nejmoa1602295.
5. Greenwood S, Perrin A, Duggan M. Social media update 2016 [Internet]. Washington, DC: Pew Research Center; 2016 Nov 11 [cited 2016 Nov 21]. Available from: www.pewinternet.org/2016/11/11/social-media-update-2016/.
6. Twitter: Number of monthly active users 2010-2017 [Internet]. New York, NY: Statista, Inc; 2017 [cited 2018 Jan 19]. Available from: www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/.
7. Faasse K, Chatman CJ, Martin LR. A comparison of language use in pro- and anti-vaccination comments in response to a high profile Facebook post. Vaccine 2016 Nov 11;34(47):5808-14. DOI: https://doi.org/10.1016/j.vaccine.2016.09.029.
8. Guidry JP, Carlyle K, Messner M, Jin Y. On pins and needles: How vaccines are portrayed on Pinterest. Vaccine 2015 Sep 22;33(39):5051-6. DOI: https://doi.org/10.1016/j.vaccine.2015.08.064.
9. Larson HJ, Wilson R, Hanley S, Parys A, Paterson P. Tracking the global spread of vaccine sentiments: The global response to Japan's suspension of its HPV vaccine recommendation. Hum Vaccin Immunother 2014;10(9):2543-50. DOI: https://doi.org/10.4161/21645515.2014.969618.
10. Bessi A, Zollo F, Del Vicario M, Scala A, Caldarelli G, Quattrociocchi W. Trend of narratives in the age of misinformation. PloS One 2015 Aug 14;10(8):e0134641. DOI: https://doi.org/10.1371/journal.pone.0134641.
11. Del Vicario M, Bessi A, Zollo F, et al. The spreading of misinformation online. Proc Natl Acad Sci U S A 2016 Jan 19;113(3):554-9. DOI: https://doi.org/10.1073/pnas.1517441113.
12. Bahk CY, Cumming M, Paushter L, Madoff LC, Thomson A, Brownstein JS. Publicly available online tool facilitates real-time monitoring of vaccine conversations and sentiments. Health Aff (Millwood) 2016 Feb;35(2):341-7. DOI: https://doi.org/10.1377/hlthaff.2015.1092.
13. Massey PM, Leader A, Yom-Tov E, Budenz A, Fisher K, Klassen AC. Applying multiple data collection tools to quantify human papillomavirus vaccine communication on Twitter. J Med Internet Res 2016 Dec 5;18(12):e318. DOI: https://doi.org/10.2196/jmir.6670.
14. Zhou X, Coiera E, Tsafnat G, Arachi D, Ong MS, Dunn AG. Using social connection information to improve opinion mining: Identifying negative sentiment about HPV vaccines on Twitter. Stud Health Technol Inform 2015;216:761-5. DOI: https://doi.org/10.3233/978-1-61499-564-7-761.
15. Suh B, Hong L, Pirolli P, Chi EH. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. Proceedings of the 2010 IEEE Second International Conference on Social Computing; 2010 Aug 20-22; Minneapolis, MN. New York, NY: IEEE; 2010 Sep 30. DOI: https://doi.org/10.1109/socialcom.2010.33.
16. Infographics: #VaccinesWork [Internet]. Geneva, Switzerland: World Health Organization; 2017 Apr [cited 2017 May 27]. Available from: www.who.int/campaigns/immunization-week/2017/infographic/en/.
17. Rodríguez G. Models for count data with overdispersion [Internet]. Princeton, NJ: Princeton University; 2013 Nov 6 [cited 2017 Mar 22]. Available from: http://data.princeton.edu/wws509/notes/c4a.pdf.
18. Talking to parents about vaccines [Internet]. Atlanta, GA: Centers for Disease Control and Prevention; 2015 Nov 30 [cited 2016 Nov 20]. Available from: www.cdc.gov/vaccines/hcp/conversations/conv-materials.html.
19. Shearer E, Gottfried J. News use across social media platforms 2017 [Internet]. Washington, DC: Pew Research Center; 2017 Sep 7 [cited 2017 Dec 10]. Available from: www.journalism.org/2017/09/07/news-use-across-social-media-platforms-2017/.
20. Bessi A, Caldarelli G, Del Vicario M, Scala A, Quattrociocchi W. Social determinants of content selection in the age of (mis)information. In: Aiello LM, McFarland D, editors. Social informatics. Proceedings of SocInfo 2014, the 6th International Conference on Social Informatics; Barcelona, Spain; 2014 Nov 11-13. Cham, Switzerland: Springer International Publishing AG; 2014. p 259-68. DOI: https://doi.org/10.1007/978-3-319-13734-6.
21. Kang GJ, Ewing-Nelson SR, Mackey L, et al. Semantic network analysis of vaccine sentiment in online social media. Vaccine 2017 Jun 22;35(29):3621-38. DOI: https://doi.org/10.1016/j.vaccine.2017.05.052.
22. Fung IC, Hao Y, Cai J, et al. Chinese social media reaction to information about 42 notifiable infectious diseases. PLoS One 2015 May 6;10(5):e0126092. DOI: https://doi.org/10.1371/journal.pone.0126092. Erratum in: PLoS One 2015 May 20;10(5):e0129525. DOI: https://doi.org/10.1371/journal.pone.0129525.
23. Fung ICH, Zeng J, Chan CH, et al. Twitter and Middle East respiratory syndrome, South Korea, 2015: A multi-lingual study. Infect Dis Health 2018 Mar;23(1):10-6. DOI: https://doi.org/10.1016/j.idh.2017.08.005.