Int Health
doi:10.1093/inthealth/ihy087

World Pneumonia Day 2011–2016: Twitter contents and retweets

Md Mohiuddin Adnan[a,†], Jingjing Yin[a,†], Ashley M. Jackson[a], Zion Tsz Ho Tse[b], Hai Liang[c], King-Wa Fu[d,e], Nitin Saroha[f], Benjamin M. Althouse[g,h,i] and Isaac Chun-Hai Fung[a,*]

[a]Jiann-Ping Hsu College of Public Health, Georgia Southern University, Statesboro, GA 30460, USA; [b]School of Electrical and Computer Engineering, College of Engineering, University of Georgia, Athens, GA 30602, USA; [c]School of Journalism and Communication, Chinese University of Hong Kong, Hong Kong; [d]Journalism and Media Studies Centre, University of Hong Kong, Hong Kong; [e]MIT Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; [f]Department of Computer Science, University of Georgia, Athens, GA 30602, USA; [g]Institute for Disease Modeling, Bellevue, WA 98005, USA; [h]Information School, University of Washington, Seattle, WA 98105, USA; [i]New Mexico State University, Las Cruses, NM 88003, USA

*Corresponding author: Tel: +1 912 4785079; Fax: +1 912 4780171; E-mail: cfung@georgiasouthern.edu
[†]Co-first authors.

Received 27 April 2018; revised 4 September 2018; editorial decision 3 October 2018; accepted 17 November 2018

Background: Twitter is used for World Pneumonia Day (WPD; November 12) communication. We evaluate if themes of #pneumonia tweets were associated with retweet frequency.

Methods: A total of 28 181 original #pneumonia tweets were retrieved (21 November 2016), from which six subcorpora, 1 mo before and 1 mo after WPD 2011–2016, were extracted (n=6721). Underlying topics were identified via latent Dirichlet allocation and were manually coded into themes. The association of themes with retweet count was assessed via multivariable hurdle regression.

Results: Compared with personal experience tweets, tweets that both raised awareness and promoted intervention were 2.62 times as likely to be retweeted (adjusted odds ratio [aOR] 2.62 [95% 1.79 to 3.85]) and if retweeted had 37% more retweets (adjusted prevalence ratio [aPR] 1.37 [95% CI 1.06 to 1.78]). Tweets that raised concerns about vaccine price were twice as likely to be retweeted (aOR 2.29 [95% CI 1.36 to 3.84]) and if retweeted, had double the retweet count (aPR 2.05 [95% CI 1.27 to 3.29]) of tweets sharing personal experience.

Conclusions: The #pneumonia tweets that both raised awareness and promoted interventions and those discussing vaccine price were more likely to engage users than tweets about personal experience. These results help health professionals craft WPD messages that will engage the audience.

Keywords: health communication, health marketing, machine learning, manual coding, social media

Introduction

Pneumonia was estimated to cause around 935 000 deaths in children <5 y of age worldwide in 2000–2013.[1] Globally, 2.7 million deaths of all age groups were attributable to lower respiratory infections in 2015.[2]

World Pneumonia Day (WPD) was established on 12 November 2009 to raise awareness about pneumonia, to promote effective interventions and to encourage individuals and organizations worldwide to take actions against it.[3] Since 2009, WPD has been a global event through which public health and medical organizations galvanize public support for pneumonia research and promote immunization and other interventions against pneumonia.[3] According to the 2011 WPD report, WPD has three goals: 'to raise awareness about the disease' (raising awareness), to 'promote interventions to protect against, prevent and treat pneumonia as called for in the Global Action Plan for the Prevention and Control of Pneumonia' (promoting intervention) and to 'generate action to combat the world's leading killer of young children' (call to action).[3]

Social media has become a vital channel for health organizations to communicate with the public. Twenty-four percent of US adults post about personal health experiences whereas 16% of them post reviews of treatments, doctors and medication online.[4] Understanding and interpreting topics and themes of online posts pertinent to a health topic will assist health communicators to better understand the 'communication environment' in which they operate. Health communicators use social media to promote hand hygiene,[5] to promote vaccination and understand vaccine hesitancy,[6] and to understand tobacco marketing and advocate for tobacco control, among others.[7] Likewise, risk communication can be conducted via social media as public health agencies respond to natural disasters[8] and infectious disease outbreaks (e.g. Ebola,[9] Zika[10] and Middle East respiratory syndrome).[11] A recent study of US federal health agencies' engagement on Facebook demonstrates that online networking records of various health agencies differ in their use of social media and their engagement with social media users.[12] According to the same study, health agencies post more on Twitter than other social media platforms such as Facebook.[12] For example, the Centers for Disease Control and Prevention (CDC) uses Twitter to host Twitter chats on Ebola and Zika[13,14] and to advertise their monthly events and publications.[15] With about 328 million monthly active users in 2017, Twitter is one of the most popular social media platforms worldwide.[16] Twitter provides a communication channel through which laypeople can advocate for science, as seen in tweets pertinent to climate change.[17] Thus analysis of Twitter content and retweet frequency can provide public health professionals an intermediary measure of the reach or viability of a given health message or general health campaign.[18] Ultimately the goal of enhancing the performance of social media marketing of health campaigns is to reduce disease burden through reaching a larger audience with health messages and increasing the adoption of healthy behaviors.

The objective of this study was to summarize and interpret the contents of tweets with #pneumonia around WPD 2011–2016. To assist health communicators in expanding the reach of their tweets, our specific research question was to identify the most popular contents (which were integrated into themes) for #pneumonia tweets, measured in terms of retweet frequency. In particular, we investigated if tweets mentioned topics that fell in line with the three goals described in the 2011 WPD report.[3]

Methods

The entire corpus of original tweets with #pneumonia (case insensitive) from 16 September 2011 through 21 November 2016 was retrieved via the Twitter Search Application Programming Interface. We extracted those tweets that were tweeted within 1 mo before and 1 mo after WPD—from 13 October to 12 December—for the first 5 y (2011–2015). Since our data were retrieved on 21 November 2016 for 2016, we extracted those tweets that were tweeted from 13 October to 21 November. These six subsets (subcorpora) of Twitter data, a total of 6721 tweets, were used for subsequent analysis. In the following sections, these 6721 tweets were referred to as the 'entire sample', to be distinguished from the 'entire corpus' of original #pneumonia tweets (n=28 181).

Data analysis was conducted in R version 3.3.1 (R Project for Statistical Computing, Vienna, Austria). To categorize the contents of the tweets, we used an unsupervised machine learning method, latent Dirichlet allocation (LDA), which is a type of probabilistic topic model.[19] The LDA model automatically assigns the probabilities of being in each of the topics by looking at the individual term's joint distributions from each of the tweets. Prior to topic modeling, we deleted the keyword 'pneumonia' (as it appeared in all tweets in our dataset), years, Uniform Resource Locator (URL) links, hashtags and stop words from each tweet and created a document–term matrix. Since the LDA model needed a prespecified value for the number of topics, for each subcorpus we tested models with numbers of topics from 5 to 100, in increments of 5, each with 50 replications. From the perplexity plot, the optimal number of topics was chosen as the value that yielded a global minimum of perplexity score (Supplementary Table 1; Supplementary Figures 1–6). We applied the LDA model with the corresponding optimal number of topics for the document–term matrix of each subcorpus to allocate each tweet to the most probable topic. We manually named each topic by reading a few example tweets with the top probabilities assigned to that topic. The first co-first author categorized those themes into three a priori designated binary themes based on the three goals of WPD as stated in the 2011 report: 'to raise awareness about the disease' (raising awareness), 'promote interventions to protect against, prevent and treat pneumonia as called for in the Global Action Plan for the Prevention and Control of Pneumonia' (promoting intervention) and 'generate action to combat the world's leading killer of young children' (call to action).[3] Some topics were assigned to more than one theme, whereas some topics did not have information about any of these three themes. Also, during the process of manual coding of topics, we noticed that many tweets were about individual experiences of pneumonia and these tweets did not fall into the prescribed categories. Thus one additional theme was created: personal experience. If a topic did not fall into any of these themes, it was categorized as miscellaneous. A total of 532 tweets in the entire sample of 6721 tweets were categorized as miscellaneous (7.91%). However, we noted about one in five tweets were categorized as miscellaneous in the subcorpora of 2015 (284/1221 [23.26%]) and 2016 (134/688 [19.48%]). For this reason, M.M.A went through each of the topics as well as the tweets to identify potential additional themes from the miscellaneous topics of tweets. Based on these topics, the authors agreed to create three additional themes for 2015—price of vaccine, Pneumonia Innovations Summit and invitation to learn more from other sources—and one additional theme for 2016—price of vaccine. We then back-assigned each tweet with the themes to which its topic was assigned (Table 1). Because categorizing from topics to the theme was a manual process, the corresponding author independently coded a random 10% sample of the topics of each subcorpus. The Cohen's unweighted κ was 0.79, representing substantial interrater reliability.

To assess the association between retweet frequency (the dependent variable) and themes, after controlling for potential confounders, multivariable regression models were applied to each subcorpus as well as to a combined data set of the six subcorpora (i.e. the 'entire sample') to obtain overall estimates over the 6 consecutive years. Hurdle regression models were chosen to account for excesses of zero frequency in the dependent variable (i.e. the number of retweets). A hurdle regression model encompasses two models, a logistic regression model for modeling the presence or absence of retweets (zero-hurdle model right censored at y=1) and a truncated count model for retweet counts if retweet does happen (count data model left-truncated at y=1).[20] Negative binomial distribution was assumed for the

Table 1. Inclusion criteria and example tweet for each content theme for a sample of #pneumonia tweets from 2011 to 2016

Theme name	Inclusion criteria	Example tweet
Raising awareness	Tweets that had information, statistics, global facts, awareness, key messages, etc. about pneumonia	*Did you know 98% of children who die from #Pneumonia live in developing countries #WPD2011 #Egypt #EPSF #IPSF #WHO* http://t.co/colY5lH7
Promoting the intervention	Tweets that had an indication of any kind of preventive and/or treatment method (e.g. vaccine, hand washing, breastfeeding, hospitalization)	*#Vaccines breastfeeding nutrition handwashing & reducing indoor air pollution help prevent #pneumonia #wpd2011*
Call to action	Tweets that had an indication of urging help from governments, political leaders, health professionals, researchers, individuals with the aim of preventing and treating pneumonia	*Show support for #wpd2011 donate #10 to @GAVIAlliance to provide 1 child w lifetime of protection against #pneumonia* http://t.co/3yZMLHq4
Personal experience	Tweets that had information about the individual concern of having pneumonia, bad weather, etc.	*Freezing wet bike ride to my 8am in the dark. #frozenhands #pneumonia #puremichigan*
Price of vaccine	Tweets about the high price of vaccines and the concern that most children will not be covered; why Médecins Sans Frontières (MSF), USA rejected donation on vaccine from Pfizer, etc.	*Why is #Pneumonia vaccine out of reach for many children? Sign our petition to @Pfizer &amp; @GSK:* https://t.co/sAi0PyLYxR
Pneumonia Innovations Summit	Tweets about campaign for Pneumonia Innovations Summit (2015)	*World Pneumonia Day is on November 12. Vote for your favorite #pneumonia #innovator today!* https://t.co/N3f2Uue8bD https://t.co/KAN1UTeRDP
Inviting to learn more from other sources	Tweets asking for reading, watching materials to learn more about pneumonia, to join in a discussion or webinar or to watch PowerPoint slides etc.	*Print and handout this informative pamphlet about #pneumonia to your patients:* https://t.co/Far1EgRG8Q
Raising awareness and promoting intervention	Tweets about raising awareness; fighting pneumonia; knowing the facts, symptoms, preventive methods; campaign of vaccination; hospitalization; antibiotics etc.	*Today is #worldpneumoniaday! Did you know regular #handwashing with soap is an effective way to prevent #pneumonia?* https://t.co/Jf8BWgH7Id
Raising awareness and call to action	Tweets about raising awareness, urging help to spread facts and different kinds of information about pneumonia; urging support to fight pneumonia	*Help raise awareness of the leading cause of death in children under 5 worldwide - #Pneumonia. Join the campaign at* worldpneumoniaday.org

truncated count model given the overdispersion of the count data. The confounders chosen to be included in our regression models include (a) user characteristics (numbers of followers, friends [Twitter users whom one follows], favorites [tweets one 'likes'] and status updates [the total number of tweets one ever tweeted]) and (b) tweet-specific meta-data (the age of a tweet, hashtag count and the presence of a URL link).[15,21] In the model for the 'entire sample' we also included the day of the week, based on prior literature on the circaseptan rhythm of health information seeking as found on Google Search.[22] Thursday was chosen as the reference category because the first tweet of our sample was posted on a Thursday (13 October 2011). The value of 0.05 was chosen as the level of significance.

Finally, correspondence analysis was performed for the top 10 topics of each subcorpus (Supplementary Tables 4, 7–11) to explore the relationships between themes and tweet and retweet counts in different years (Supplementary Tables 12 and 13).

Results

A total of 28 181 original tweets with #pneumonia were retrieved from 16 September 2011 through 21 November 2016. The entire corpus received 46 709 retweets over 6 y, from 2011 through 2016 (63 mo). As WPD happens every November, the frequencies of both tweets and retweets were generally greater in November than any of the other months during that period of 63 mo (see Figure 1). The year of 2016 was an exception, in which (a) the highest numbers of original tweets and their retweets were in September and those tweets were about one of the US presidential candidates getting pneumonia and (b) the data were censored on the date of data retrieval (i.e. 21 November). The overall pattern of the retweet count shows that the number of retweets of pneumonia-related tweets increased gradually over time (Figure 1). Of all original tweets, 27 545 (97.74 %) were in English.

In our sample, the original tweet frequencies of the top 10 users with the highest number of tweets were 316, 237, 183, 169, 167, 138, 118, 114, 114 and 103, respectively. However, the total number of retweets of the tweets written by these users varied from 0 to more than 1000: 2, 608, 71, 4, 0, 1097, 1074, 633, 142 and 75, respectively (descriptions of the top 10 users are in Supplementary Table 2). Interestingly, the top 10 retweeted tweets were not generated by any of the top 10 users (ranked by the number of retweets of a single tweet). The top 10 most retweeted tweets (retweet frequencies of those tweets were 1261, 467, 415, 292, 278, 270, 261, 255, 225 and 220) were written by four users: UNICEF wrote seven and the

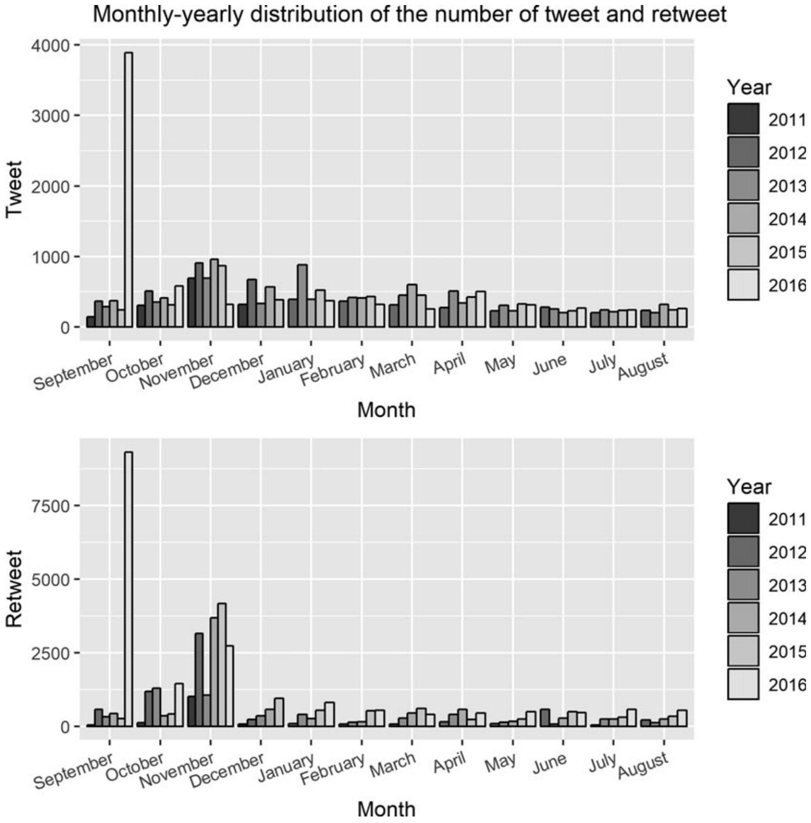

Figure 1. Monthly frequency of #pneumonia tweets (upper panel) and their retweets (lower panel) over 6 y (from 16 September 2011 through 21 November 2016). Note that the peak of original tweets in 2016 was in September, not November. In September 2016, the news of politician Hillary Clinton suffering from pneumonia attracted a lot of attention from Twitter users. Thus September 2016 instead of November had the greatest number of original tweets and retweets. In November 2016 (until 21 November), there were 321 original tweets.

WHO and celebrities Melissa Joan Hart and Mandy Moore each wrote one (Supplementary Table 3).

The percentage of tweets that were retweeted at least once increased from 24.34% to 57.00% between 2011 and 2015 (Table 2). In 2016, the percentage was smaller than in 2015 but greater than in any of the previous years (2011–2014).

For the topic model, the best model (the model with the lowest perplexity score) for each year's sample contains the following number of topics: 45 (2011), 50 (2012), 55 (2013), 20 (2014), 35 (2015) and 35 (2016) (Supplementary Table 1; Supplementary Figures 1–6). An example tweet of each content theme is provided in Table 1. The percentage of tweets falling into different themes varied: raising awareness (range 19.48–54.93%), promoting intervention (range 8.59–31.83%), call to action (range 0–4.65%), personal experience (range 17.94–51.31%) and miscellaneous (range 0–8.87%) (Table 2). Figure 2 presents the numbers of original tweets in the four main themes over 6 y. No topics were categorized as calls to action in 2012, 2013 or 2014 and there were no miscellaneous topics in 2013 or 2014 (Table 2).

After adjusting for confounders in a multivariable hurdle model, the relative risks of retweets are presented in Table 3. Compared with tweets sharing personal experiences, tweets that voiced concern about the price of vaccines were 2.29 times (95% 1.36 to 3.84) as likely to be retweeted, and if retweeted had 2.05 times (95% CI 1.27 to 3.29) the number of retweets. Tweets that both raised awareness and promoted interventions were 2.62 times (95% CI 1.79 to 3.85) as likely to be retweeted, and if retweeted had 1.37 times (95% CI 1.06 to 1.78) as many retweets as tweets sharing one's personal experience.

Compared with tweets sharing personal experiences, if retweeted, tweets that only raised awareness had 79% more retweets (adjusted prevalence ratio [aPR] 1.79 [95% CI 1.52 to 2.11]), while tweets that only promoted intervention had 31% more retweets (aPR 1.31 [95% CI 1.08 to 1.58]), tweets calling users to action had 144% more (aPR 2.44 [95% CI 1.30 to 4.59]) and tweets that both raised awareness and called users to actions had 132% more (aPR 2.32 [95% CI 1.35 to 3.99]).

#Pneumonia tweets in recent years received more retweets. After adjusting for any secular trends in retweeting over time, an increase in 1 y of age of a tweet reduced its probability of being retweeted by 17% (aOR 0.83 [95% CI 0.78 to 0.87]), and if the tweet was retweeted, an increase in 1 y of age caused it to receive 19% fewer retweets (aPR 0.81 [95% CI 0.77 to 0.84]).

The user matters. If the Twitter user who tweeted the tweet had 10 times the number of followers, the tweet was 3.83 times (95% CI 3.47 to 4.23) as likely to be retweeted, and if retweeted it had 4.54 times (95% CI 4.10 to 5.04) the number of retweets.

Table 2. Number of tweets (and percentage) in samples of #pneumonia tweets around WPD (13 October–12 December), 2011–2016, by the number of retweets and content themes

	Original tweets in each sample, n (%)					
Year of sample	2011	2012	2013	2014	2015	2016[a]
Total, n	998	1409	1036	1369	1221	688
By number of retweets						
Retweeted at least once	243 (24.34)	451 (32.00)	360 (34.75)	566 (41.34)	696 (57.00)	303 (44.04)
Never retweeted	755 (75.66)	958 (68.00)	676 (65.25)	803 (58.66)	525 (43.00)	385 (55.96)
By content theme						
Raising awareness	306 (30.66)	395 (28.03)	352 (33.97)	752 (54.93)	345 (28.26)	134 (19.48)
Promoting intervention	115 (11.52)	121 (8.59)	198 (19.11)	247 (18.04)	284 (23.26)	219 (31.83)
Call to action	11 (1.10)	0 (0)	0 (0)	0 (0)	14 (1.15)	32 (4.65)
Price of vaccine	0 (0)	0 (0)	0 (0)	0 (0)	54 (4.42)	46 (6.69)
Pneumonia Innovations Summit (2015)	0 (0)	0 (0)	0 (0)	0 (0)	36 (2.95)	0 (0)
Inviting to learn more from other sources	0 (0)	0 (0)	0 (0)	0 (0)	94 (7.69)	32 (4.65)
Raising awareness and promoting intervention	83 (8.32)	69 (4.89)	31 (2.99)	123 (8.98)	96 (7.86)	13 (1.89)
Raising awareness and call to action	56 (5.61)	16 (1.13)	4 (0.39)	0 (0)	0 (0)	0 (0)
Personal experience	398 (39.88)	723 (51.31)	451 (43.53)	247 (18.04)	219 (17.94)	151 (21.95)
Miscellaneous	29 (2.90)	85 (6.03)	0 (0)	0 (0)	79 (6.47)	61 (8.87)

[a] This sample is censored at 21 November 2016, the date of data retrieval.

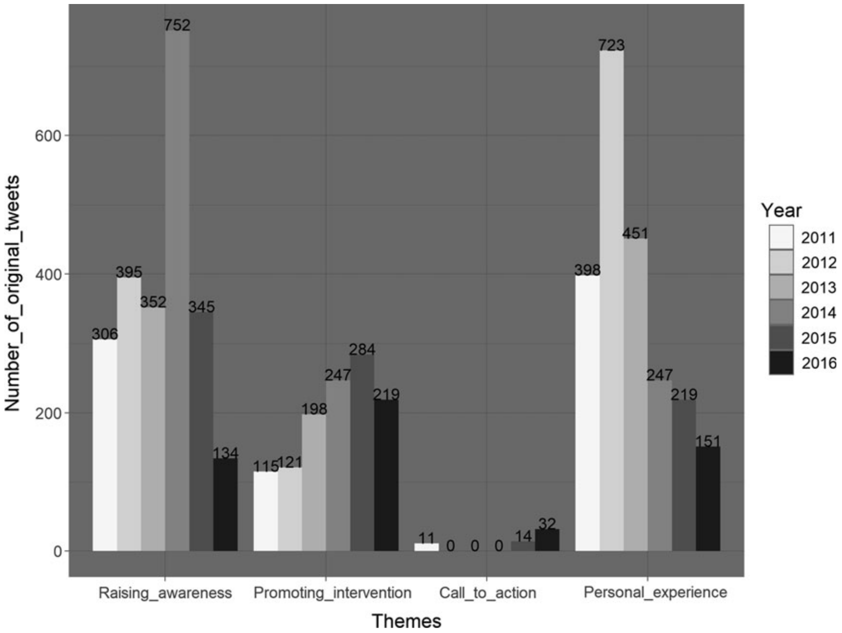

Figure 2. Number of original tweets for four main themes over 6 y (2011–2016).

If the Twitter user who tweeted the tweet followed 10 times more users (an increased friend count by 10 times), his/her tweets had 13% fewer retweets (aPR 0.87 [95% CI 0.77 to 0.98]).

If the total number of tweets that the Twitter user ever tweeted (the number of status updates) increased by 10-fold, the tweet was 36% less likely to be retweeted (aOR 0.64 [95% CI 0.56 to 0.74]), and if retweeted there were 60% fewer retweets (aPR 0.40 [95% CI 0.36 to 0.45]) (Table 3). If the Twitter user had 10 times more favorites (i.e. to 'like' a tweet), his/her tweet would have 19% (aPR 1.19 [95% CI 1.12 to 1.28]) more retweets.

While lower retweet probability was observed for Friday, Sunday and Tuesday, the day of the week when the tweet was tweeted did not significantly affect retweet probability.

The visualization of the correspondence analysis (Figure 3; Supplementary Table 12) shows that some themes (indicated

Table 3. OR estimates, 95% CIs and p-values for being retweeted vs not being retweeted (aOR from zero-hurdle component of the hurdle model) and corresponding values for positive retweet counts (aPR from negative binomial regression model component of the hurdle model) for the 'entire sample' of 6721 tweets

Coefficients	aOR of being retweeted vs not being retweeted (95% CI)	p-Value	Adjusted prevalence ratio of retweet count (95% CI)	p-Value
Content themes				
Personal experience	Reference	–	Reference	–
Raising awareness	1.25 (1.00 to 1.57)	0.05	1.79 (1.52 to 2.11)	<0.01
Promoting intervention	0.87 (0.68 to 1.13)	0.30	1.31 (1.08 to 1.58)	<0.01
Call to action	1.92 (0.98 to 3.77)	0.06	2.44 (1.30 to 4.59)	<0.01
Price of vaccine	2.29 (1.36 to 3.84)	<0.01	2.05 (1.27 to 3.29)	<0.01
Pneumonia Innovations Summit	0.55 (0.22 to 1.37)	0.20	0.71 (0.34 to 1.48)	0.36
Inviting to learn more from other sources	0.64 (0.38 to 1.08)	0.10	1.46 (0.95 to 2.24)	0.08
Raising awareness and promoting intervention	2.62 (1.79 to 3.85)	<0.01	1.37 (1.06 to 1.78)	0.02
Raising awareness and call to action	1.39 (0.72 to 2.70)	0.33	2.32 (1.35 to 3.99)	<0.01
Miscellaneous	1.00 (0.64 to 1.56)	0.99	1.22 (0.88 to 1.70)	0.24
Confounders				
Hashtags (count)	0.96 (0.90 to 1.02)	0.15	1.03 (0.99 to 1.08)	0.11
Tweet age (y)	0.83 (0.78 to 0.87)	<0.01	0.81 (0.77 to 0.84)	<0.01
URL link in tweet body (binary)	1.24 (1.04 to 1.47)	0.01	0.82 (0.71 to 0.94)	<0.01
Follower count (log 10)	3.83 (3.47 to 4.23)	<0.01	4.54 (4.10 to 5.04)	<0.01
Friend count (log 10)	0.96 (0.83 to 1.11)	0.60	0.87 (0.77 to 0.98)	0.02
Favorite count (log 10)	0.91 (0.83 to 1.01)	0.10	1.19 (1.12 to 1.28)	<0.01
Status update count (log 10)	0.64 (0.56 to 0.74)	<0.01	0.40 (0.36 to 0.45)	<0.01
Thursday	Reference	–	Reference	–
Friday	0.65 (0.51 to 0.84)	<0.01	0.90 (0.74 to 1.09)	0.28
Saturday	1.09 (0.81 to 1.46)	0.58	0.82 (0.65 to 1.02)	0.08
Sunday	0.68 (0.48 to 0.97)	0.03	0.83 (0.64 to 1.08)	0.17
Monday	0.78 (0.60 to 1.02)	0.07	0.97 (0.79 to 1.19)	0.78
Tuesday	0.75 (0.58 to 0.97)	0.03	0.80 (0.66 to 0.98)	0.03
Wednesday	0.95 (0.75 to 1.22)	0.69	0.99 (0.82 to 1.20)	0.94

The primary independent variables in our regression model are the themes of tweets. Potential confounders were included, including hashtag count; tweet age (in years); the presence or absence of a URL link in the tweet body; the users' follower count, friend count, favorite count and status update count (number of tweets ever tweeted); and the day of the week (with Thursday as the reference category). It is important to note that contents of certain themes might not exist in the subcorpora of all years under study. In addition, we added 1 to each observation for followers, favorites, friends and status to avoid the problem with log transformation of zero. The log-likelihood of the model was −9015 (with degrees of freedom=47).

by points) were more discriminating by the year (indicated by triangles) whereas some were not. For example: the themes 'call to action', 'price of vaccine', 'personal experience' and 'raising awareness and call to action' are highly separated by the first two principal components, which retain 60.3% of the total variability of the data, while 'raising awareness', 'promoting the intervention' and 'raising awareness and promoting intervention' are dominant themes across all years. Furthermore, the themes 'personal experience' and 'raising awareness and call to action' were posted more often during 2011, 2012 and 2013, but the themes 'raising awareness' and 'raising awareness and promoting awareness' were retweeted more during those years. Tweets having the themes 'promoting the intervention', 'invitation of learning more from other sources', 'call to action' and 'price of vaccine' were retweeted more in 2015 and 2016.

Discussion

We outlined descriptive statistics and computational content analysis of pneumonia-related tweets around WPD from 2011 to 2016. We found that compared with tweets sharing one's personal experiences of pneumonia, pneumonia-related tweets that both raised awareness and promoted interventions and that mentioned vaccine price were more likely to be retweeted (p<0.01 in both cases). Given that such tweets attracted heightened attention from and the engagement of Twitter users, global health advocates can attempt to raise awareness and promote interventions in the same tweet.

The negative association between retweet probability and tweet age (in years) highlighted that there was a trend of increasing retweet frequency over time for #pneumonia tweets.

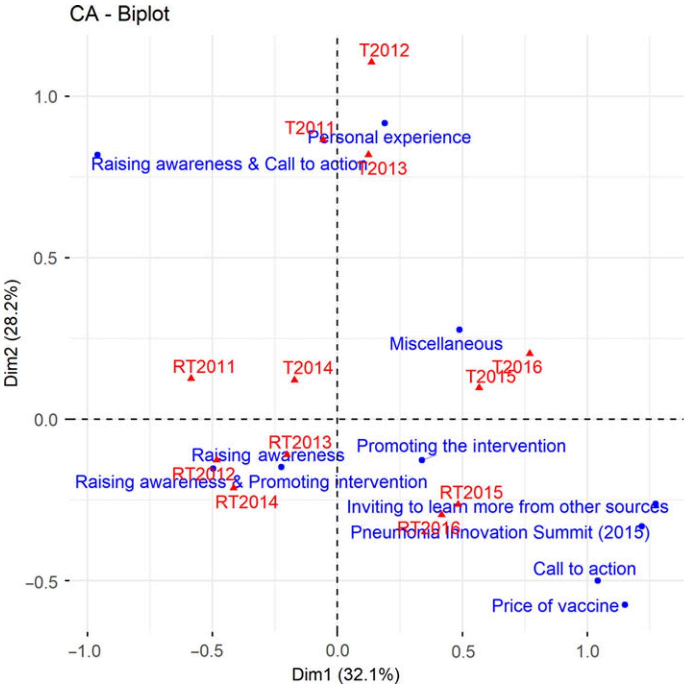

Figure 3. Correspondence analysis of the themes of the tweets in the top 10 topics and their retweets from 2011 to 2016.

This is likely to be true for all tweets in general. Among a stratified random sample of 2126 users, the median retweet time of the tweets of half of the users was ≤18 min.[23] It is extremely rare to have a tweet still being retweeted a year later. Therefore the negative association between retweet probability and tweet age should not be interpreted as a year-old tweet still having some probability of being retweeted today. Instead, the result should be interpreted in light of the ever-growing number of Twitter users globally. It reflects that a #pneumonia tweet posted today is more likely to be retweeted than was a #pneumonia tweet tweeted last year. People became more engaged with Twitter for posting and sharing pneumonia-related information each year. The retweet frequency increased greatly in the time surrounding WPD each year, which suggests that people responded to this health communication event on Twitter. Our findings demonstrate that pneumonia-related Twitter health promotion around the time of WPD triggered reactions among Twitter users. Therefore health communication professionals should continue their practice of promoting pneumonia awareness and prevention on Twitter yearly at the time of WPD.

The users who tweeted most about #pneumonia were not necessarily the users whose tweets were retweeted most. The most retweeted tweets were those posted by health organizations. The user @ExpatInc, who tweeted 316 original tweets with a URL link to promote a website, was probably a ‘bot’. Nevertheless, a recent study suggests that a health-related tweet posted by a non-celebrity Twitter user can also go viral (with a high retweet count) depending on its content.[24] This observation might suggest that the content and/or users of tweets impacted the retweet count. As our study’s purpose was to analyze the contents of the tweets rather than meta-data of Twitter accounts, we investigated tweets coded with a single theme and those in combinations of themes. We found that tweets that both raised awareness and promoted intervention were more likely to be retweeted and to have more retweets than tweets sharing one’s personal experiences. In 2015 and 2016, the price of and access to vaccines was found to be the theme of a number of #pneumonia tweets. As vaccine hesitancy is on the rise, as more than 50 vaccines are being developed to prevent respiratory syncytial virus,[25] more could be done with Twitter health communication by engaging users with information pertinent to the price, access, effectiveness and other relevant aspects of vaccines. Even though appealing for help from governments, health professionals and individuals was one of the main purposes of WPD,[3] only a small percentage of tweets conveyed ‘call to action’ messages.

Having more followers is associated with more retweets; however, tweeting too many tweets is associated with a lower retweet probability. This is in line with the existing literature.[26] We need timely tweets from users with a large number of followers, but without posting too many tweets as to overwhelm the followers with information.

Using unsupervised machine learning to uncover the underlying topics of thousands of tweets coupled with the manual coding of scores of topics into several major themes appears to be a feasible method that public health practitioners can use in practice. It strikes a balance between time efficiency and data interpretability. Our prior research applied the same method to analyze Twitter users’ responses to the WHO’s declaration of Zika virus as a public health emergency of international concern,[10] as well as disease-specific contents on malaria, human immunodeficiency virus, tuberculosis, noncommunicable diseases and neglected tropical diseases in the context of Twitter #globalhealth conversations.[27] This study further demonstrates the feasibility of this approach. This is good news to many global health advocates, as their social media health communication efforts can be quantitatively measured and empirically evaluated, at least at the level of social media posts.

This study is subject to certain limitations. While categorizing tweets into topics for different subcorpora by using an unsupervised machine learning method speeded up the content analysis process, potential misclassification was possible. The manual process of naming of topics and categorization of those topics into themes requires human judgement and is prone to human errors. Nonetheless, our interrater reliability was substantial. Additionally, the semiautomated analysis combined quantitative and qualitative analysis and thus had the advantages of both. Our corpus was retrieved using a hashtag (#pneumonia) in the English language and thus most tweets in our corpus were in English. We did not investigate the geolocations of the users. While celebrities might have an impact on the retweet frequency of a health topic,[28] as we found here with Melissa Joan Hart and Mandy Moore having 2 of the top 10 retweeted tweets, an analysis of users’ profiles is beyond the scope of this study and further research is necessary.

To conclude, along with the increasing number of Twitter users, public engagement in obtaining and sharing information about pneumonia on Twitter (as measured by retweet frequency) increased from 2011 to 2016. Furthermore, in recent years people have been concerned about the price of and access to vaccines, a challenge that is also the concern of many public health professionals.[29] Twitter provides a platform through which public health professionals can raise awareness, provide evidence-based information and mobilize supporters to take action pertinent to global health issues such as pneumonia. Specifically, tweets from users with large numbers of followers may need to focus on messages emphasizing awareness of and promoting interventions against pneumonia. The policy implication of this study is that scientific message design and testing to increase social media engagement should be part of health marketing strategies adopted for WPD or other similar events.

Supplementary data

Supplementary data are available at International Health online (http://inthealth.oxfordjournals.org).

Authors' contributions: MMA, JY and ICHF conceived and designed the study. ZTHT and NS retrieved the Twitter data. MMA performed the statistical analysis under the supervision of JY and ICHF. MMA performed the manual coding, with ICHF as the second coder. AMJ assisted MMA in both writing and programming in the early phase of this project. ZTHT, HL, KWF and BMA provided MMA, JY and ICHF with consultation in statistics, computer science and health communication. MMA wrote the early drafts of the manuscript. JY and ICHF rewrote the manuscript with significant intellectual inputs. AMJ, ZTHT, HL, KWF and BMA edited the manuscript with intellectual inputs. MMA and JY serve as co-first authors. ICHF serves as the senior and corresponding author.

Acknowledgements: The opinions expressed in this paper are those of the authors and do not represent those of the CDC or the United States government.

Funding: This project was not supported by external funding. ICHF received salary support from the CDC (16IPA1609578). This project is not related to his CDC-supported project.

Competing interests: None declared.

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

References

1 Liu L, Oza S, Hogan D et al. Global, regional, and national causes of child mortality in 2000–13, with projections to inform post-2015 priorities: an updated systematic analysis. Lancet 2015;385(9966):430–40.

2 GBD 2015 Mortality and Causes of Death Collaborators. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. Lancet 2016;388(10053):1459–1544.

3 International Vaccine Access Center at Johns Hopkins Bloomberg School of Public Health. 2011 The Global Coalition Against Child Pneumonia. Fight Pneumonia. Save a Child. Baltimore, MD: International Vaccine Access Center at Johns Hopkins Bloomberg School of Public Health; 2012. https://stoppneumonia.org/wp-content/uploads/2012/04/World-Pneumonia-Day-2011-online-1.pdf (accessed 15 November 2018).

4 PricewaterhouseCoopers. Social media 'likes' healthcare: from marketing to social business. 2017. http://www.pwc.com/us/en/health-industries/health-research-institute/publications/health-care-social-media.html (accessed 22 August 2017).

5 Fung IC, Cai J, Hao Y et al. Global Handwashing Day 2012: a qualitative content analysis of Chinese social media reaction to a health promotion event. Western Pac Surveill Response J 2015;6(3):34–42.

6 Dredze M, Wood-Doughty Z, Quinn SC et al. Vaccine opponents' use of Twitter during the 2016 US presidential election: implications for practice and policy. Vaccine 2017;35(36):4670–2.

7 Huang J, Kornfield R, Szczypka G et al. A cross-sectional examination of marketing of electronic cigarettes on Twitter. Tob Control 2014;23 (Suppl 3):iii26–30.

8 Finch KC, Snook KR, Duke CH et al. Public health implications of social media use during natural disasters, environmental disasters, and other environmental concerns. Nat Hazards 2016;83(1):729–60.

9 Fung IC, Duke CH, Finch KC et al. Ebola virus disease and social media: a systematic review. Am J Infect Control 2016;44(12):1660–71.

10 Fu KW, Liang H, Saroha N et al. How people react to Zika virus outbreaks on Twitter? A computational content analysis. Am J Infect Control 2016;44(12):1700–2.

11 Fung IC-H, Zeng J, Chan C-H et al. Twitter and Middle East respiratory syndrome, South Korea, 2015: a multi-lingual study. Infect Dis Health 2018;23(1):10–6.

12 Bhattacharya S, Srinivasan P, Polgreen P. Social media engagement analysis of U.S. federal health agencies on Facebook. BMC Med Inform Decis Mak 2017;17(1):49.

13 Lazard AJ, Scheinfeld E, Bernhardt JM et al. Detecting themes of public concern: a text mining analysis of the Centers for Disease Control and Prevention's Ebola live Twitter chat. Am J Infect Control 2015;43(10):1109–11.

14 Glowacki EM, Lazard AJ, Wilcox GB et al. Identifying the public's concerns and the Centers for Disease Control and Prevention's reactions during a health crisis: an analysis of a Zika live Twitter chat. Am J Infect Control 2016;44(12):1709–11.

15 Jackson AM, Mullican LA, Yin J et al. #CDCGrandRounds and #VitalSigns: a Twitter analysis. Ann Global Health 2018. 84(4):710–6.

16 Statista. Number of monthly active Twitter users worldwide from 1st quarter 2010 to 2nd quarter 2017 (in millions). 2017. https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (accessed 7 September 2017).

17 Leas EC, Althouse BM, Dredze M et al. Big data sensors of organic advocacy: the case of Leonardo DiCaprio and climate change. PLoS One 2016;11(8):e0159885.

18 Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the US during the influenza A H1N1 pandemic. PLoS One 2011;6(5):e19467.

19 Blei DM. Probabilistic topic models. Commun ACM 2012;55(4):77–84.

20 Zeileis A, Kleiber C, Jackman S. Regression models for count data in R. J Stat Software 2008;27(8):1–25.

21 Soboleva A, Burton S, Mallik G et al. 'Retweet for a chance to. . .': an analysis of what triggers consumers to engage in seeded eWOM on Twitter. J Market Manag 2017;33(13–14):1120–48.

22 Ayers JW, Althouse BM, Johnson M et al. What's the healthiest day?: circaseptan (weekly) rhythms in healthy considerations. Am J Prev Med 2014;47(1):73–6.

23 Bray P. When is my tweet's prime of life? (a brief statistical interlude). 2012. https://moz.com/blog/when-is-my-tweets-prime-of-life (accessed 22 August 2017).

24 Noar SM, Leas E, Althouse BM et al. Can a selfie promote public engagement with skin cancer? Prev Med 2018;111:280–3.

25 Centers for Disease Control and Prevention. Announcement: World Pneumonia Day—November 12, 2016. MMWR Morb Mortal Wkly Rep 2016;65(44):1241.

26 Suh B, Hong L, Pirolli P et al. 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, Minneapolis, MN, 20–22 August 2010.

27 Fung IC, Jackson AM, Ahweyevu JO et al. #Globalhealth Twitter conversations on #Malaria, #HIV, #TB, #NCDS, and #NTDS: a cross-sectional analysis. Ann Glob Health 2017;83(3–4):682–90.

28 Noar SM, Althouse BM, Ayers JW et al. Cancer information seeking in the digital age: effects of Angelina Jolie's prophylactic mastectomy announcement. Med Decis Making 2015;35(1): 16–21.

29 Jadhav S, Datla M, Kreeftenberg H et al. The Developing Countries Vaccine Manufacturers' Network (DCVMN) is a critical constituency to ensure access to vaccines in developing countries. Vaccine 2008; 26(13):1611–5.