
Shi Chen¹, PhD; Qian Xu², PhD; John Buchenberger¹, MS; Arunkumar Bagavathi³, MS; Gabriel Fair³, MS; Siddharth Krishnan³, PhD; Samira Shaikh³, PhD

¹Department of Public Health Sciences, University of North Carolina Charlotte, Charlotte, United States

²School of Communications, Elon University, Elon, United States

³Department of Computer Sciences, University of North Carolina Charlotte, Charlotte, United States

Corresponding Author:

Shi Chen, PhD
Department of Public Health Sciences
University of North Carolina Charlotte
Room 358 Health Informatics and Analytics
9201 University City Boulevard
Charlotte, North Carolina, 28223
United States
Phone: 01 814 880 0738
Email: schen56@uncc.edu
Abstract

Background: CDC is among the first government agencies that utilize online social media (especially Twitter) to disseminate health-related information to general public. CDC was active during the 2016 Zika epidemic (which caused 5,168 domestic cases in the U.S.) on Twitter.

Objective: In this study, we aim to quantify the temporal variability in CDC’s tweeting pattern (as well as retweets and replies from general public) with regard to Zika case counts and identify potential discrepancy among actual Zika cases, CDC response on Twitter, and general public engagement (retweeting and replying).

Methods: Complete CDC-initiated tweets, including all corresponding retweets and replies were collected in 2016 from all 67 CDC-associated Twitter accounts. Time series of domestic Zika case counts was also retrieved. Both univariate (ARIMA model) and multivariate time series analyses (CCF, mutual Shannon information entropy, ARIMAX model, Granger test) were performed in each quarter of 2016.

Results: CDC sent out more than 84% of its Zika-related tweets in 1st quarter of 2016, where Zika case counts were low in both 50 U.S. states and territories. While Zika case counts increased dramatically in 2nd and 3rd quarters, CDC efforts on Twitter plunged. Time series of public engagement (characterized by ARIMA models) generally differed (based on ARIMA model parameters) among quarters as well as from that of original CDC tweets. Original CDC tweets and public engagement all had highest mutual information with Zika case counts in 2nd quarter. Public engagement was also substantially influenced by and usually preceded actual Zika epidemic.

Conclusions: There was substantial discrepancy among CDC tweets regarding Zika, public engagement, and actual Zika epidemic, as well as between different periods in 2016. We find that
although CDC was very effective in early warning of Zika in 1st quarter of 2016, public engagement was generally a more prominent predictor of actual Zika epidemic than CDC tweets later in the year.

**KEYWORDS**

Twitter; CDC; public engagement; time series analysis; Zika epidemic
Introduction

World Health Organization (WHO) has stated that health is one of the most fundamental human rights (WHO 2017) [1]. Online information platforms, especially online social media such as Twitter, Facebook, Instagram, and WeChat, have offered large coverage and almost instantaneous reach to the general public (Avery 2017) [2]. Consequently, online social media are desirable avenues to quickly disseminate accurate health and disease related information by health professionals in various topics (Avery 2017) [2]. For example, studies have shown that online social media have been utilized to monitor food safety and food-borne pathogen outbreak such as Escherichia coli O157 (Chapman 2014, Hartley 2014) [3,4], to develop online campaign to quit smoking in different countries and regions (U.S., Canada, and Hong Kong)) with various online social media platforms (Facebook, Twitter, and Whatsapp, Naslund 2017 [5]), to promote exercise, fitness, and healthy lifestyle (WeChat health campaign in China, He 2017 [6]; fitness campaign in New Orleans, LA, Rabarison 2017 [7]); to raise public awareness and engagement regarding air quality and pollution (Hu 2017) [8]; and to monitor public discussion of controversial topic such as antimicrobial resistance (Kendra 2015) [9].

Many government agencies and health officials (e.g., WHO and Centers for Disease Control and Prevention, CDC, as well as other local health departments) have also been adopting and utilizing online social media to disseminate information, communicate with public, and understand public opinions and concerns, especially during health emergency and crisis. Europe has developed an online media and crisis communication framework for influenza (Rossmann 2017) [10]. Twitter was among the top media source during the 2014 Ebola outbreak in Africa (cite). WHO and CDC utilized Twitter and Instagram during Zika outbreak (Guidry 2017) [11].
New York City monitored Zika, Hepatitis A, and Ebola discussion in social media and conducted risk communication with general public (Hadi 2017) [12].

Evidently, for many infectious diseases epidemic, it has been demonstrated that online discussion in social media can be an imperative indicator of actual disease severity and help health officials to more accurately evaluate time-sensitive epidemic situation when actual case counts are still being gathered (Paul 2014, Santillana 2015, Harris 2017) [13-15]. Time series analysis is a versatile and powerful modeling framework to link online discussion and reveal disease dynamics, as demonstrated by the extant research using autoregressive (AR) integrated moving-average (MA) with external variable (ARIMAX) model for Zika pandemic (Adebayo 2017) [16], cross-correlation function (CCF) between case counts and online discussion of influenza in social media (Broniatowski 2013) [17], and AR model with external variable for Zika forecasting in South America (McGough 2017) [18].

The 2016 Zika pandemic provides a great example to investigate and evaluate the CDC role and responsiveness in disseminating information to general public. Zika was a relatively new infectious disease, affected men and women, fetuses, and infants with multiple transmission routes, while general public had very little knowledge about it. In 2016, Zika caused 5,168 confirmed non-congenital cases in the 50 states in the U.S., and much higher case number in U.S. territories (Hall et al. 2018) [19]. CDC is among the first government agencies to utilize social media (e.g., Twitter, Facebook and Youtube channel). Twitter is the major social media outlet for CDC, with a total of 67 official CDC-associated Twitter accounts covering a wide variety of health- and disease-related topics within U.S. and across the globe (complete list available at https://www.cdc.gov/socialmedia/tools/Twitter.html). Former CDC director Dr. Tom Frieden was active on Twitter and hosted live Twitter chats with general public (Kass-hout 2013)
[20], including a recent one-hour live chat regarding Zika in February 2016. However, besides CDC’s prominent online presence and efforts, fake or inaccurate information regarding Zika proliferated on social media and outperformed CDC (and other legitimate sources such as WHO) by a large margin (Sharma et al. 2017) [21]. Studies have shown substantial discrepancy between public concern and CDC’s response to Zika on Twitter (Glowacki 2016, Joob 2016, Miller 2017, Stefanidis 2017) [22-25]. Another less addressed aspect is the low rate of public engagement (measured by number of retweets and replies) with regard to CDC’s information dissemination, where online social media should be an interactive platform for public engagement and interaction (Watts and Dodds 2007) [26], not just one-directional news outlet (Avery 2009, 2010, Hu 2017) [8, 27, 28]. Furthermore, currently there is no study on the temporal variability in CDC’s response to different epidemic stages of Zika for the entire year of 2016, its potential impact on public engagement, and quantification of information dissemination, as CDC did not finalize and publish full 2016 Zika case counts in entire U.S. until March 2018 in Morbidity and Mortality Weekly Report (MMWR, Hall et al. 2018) [19]. Thus, there is a substantial knowledge gap in quantifying and understanding the interaction among the hierarchy of Zika epidemic, CDC’s dynamic response on social media (Twitter), and public engagement to CDC’s effort, and the potential discrepancy among these hierarchies during different stages of Zika epidemic.

In this study, we aim to investigate CDC’s responsiveness on Twitter and corresponding public engagement during different stages of the 2016 Zika epidemic, and identify potential discrepancy among them, with time series analysis and information theory measurements. The insights gained from this study will reveal effectiveness of CDC’s efforts in disseminating information on social media, and help develop more effective online communication strategies to inform general public and combat fake information in other health-related topics.
Material and Methods

Data Collection, Pre-processing, and Standardization

We downloaded all English tweets with the keyword “zika” (and its associated clinical terms “microcephaly” and “Guillain-Barré Syndrome”) between 01/01/2016 and 12/31/2016 (their retweets and replies were also included), using the GNIP Twitter API through Data Science Initiative at UNC Charlotte. In addition, all tweets sent out by all 67 CDC associated accounts in 2016 were also retrieved. Zika case reports in the 50 states in the U.S. during entire 2016 have been retrieved from the official CDC Zika case report website (https://www.cdc.gov/zika/reporting/case-counts.html) as well as from CDC’s final report of 2016 Zika epidemic (MMWR, Hall et al. 2018).

Six time series were extracted from the original tweets pool and CDC case report website: number of original CDC tweets related to Zika (i.e., with keyword “Zika” and/or “microcephaly”, which represented government agency’s response to Zika epidemic); number of retweets and number of replies to CDC’s original tweets (both of which quantified general public engagement; retweets indicated neutral perspective towards the original tweet while replies might involve higher level of user engagement; Diga 2009 [29]); number of Zika case counts in the U.S. (50 states only). Since the dates of tweets/retweets/replies and case counts were not consistent (i.e., CDC may not tweet about Zika every day), these time series would be first standardized to weekly basis, i.e., aggregated in weekly period so each time series will have exactly the same 52 data points for further analysis and comparison. As a baseline scenario, weekly number of tweets (any topic) sent by all CDC accounts and top topics (identified by keywords) tweeted by CDC in 2016 were also computed and identified to demonstrate the relative importance of Zika from CDC’s perspective (i.e., ratio between weekly Zika-related
tweets and all tweets). This estimate will help reveal and assess CDC’s responsiveness to Zika at different stages of the epidemic.

**Univariate Time Series Analysis**

Original Zika-related tweets from CDC, corresponding retweets, replies, and Zika case time series were plotted, visualized, and examined for stationarity. As shown in result section, there was substantial temporal variability in original tweets, retweets, and replies, as well as Zika case, and none of these time series is stationary (Fig. 2a). To characterize such large temporal heterogeneity, we divided the entire year 2016 into four quarters and performed further analysis within each quarter. Furthermore, we calculated the ratio between Zika-related and all tweets from CDC.

These quarterly time series (excluding case counts, as most of the domestic Zika cases in 2016 were travel-related and generally cannot be characterized by ARIMA model) were first modeled as their respective autoregressive integrated moving-average (ARIMA) models to reveal any potential temporal characteristics (e.g., linear trend, seasonality, and/or temporal autocorrelation; equation 1 shows the form of an ARIMA model with variable $X_t$, difference term $L$, and parameters $p$, $d$, $q$). The optimal model was then chosen by minimizing the Akaike information criteria (AIC) value. Parameters ($p$, $d$, $q$) obtained from these optimal ARIMA models will be compared with multivariate model in next section.

$$1 - \sum_{i=1}^{p} \phi_i L^i (1-L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i \right) \epsilon_t, \text{Eqn. 1}$$

**Multivariate Time Series Analysis**
We then calculated correlation between two time series using cross-correlation function (CCF) at different stages (represented by four quarters in 2016) to identify and quantify potential discrepancy between Zika epidemic, CDC’s response (original tweets), and public engagement (retweets and replies). More specifically, Zika case versus original CDC tweets (reflecting CDC’s responsiveness to the disease outbreak); case versus retweets and case versus replies (reflecting public’s different levels of reaction to the disease) will be compared and their respective CCFs will be computed in different stages (four quarters in 2016). Since original CDC tweets vs retweets and vs replies were always highly correlated (Pearson’s correlation coefficient $\rho > 0.8$) and were not main focus of this study, we did not consider the interaction between original tweets vs retweets and vs replies in this study. Instead, we evaluated dynamic change of public engagement by calculating the ratio between CDC’s original Zika-related tweets and retweets/replies across different stages. We further calculated mutual information between two time series using Dirichlet-multinomial pseudocount Bayesian estimate of Shannon entropy, a more informative metric than CCF to reveal potential mutual information between two time series, and quantify how informative CDC’s tweets (and retweets and replies) were during different stages of Zika epidemic in 2016 (i.e., whether number of tweets/retweets/replies had adequate mutual information with actual Zika epidemic).

We constructed ARIMAX model for original CDC tweets, retweets, and replies, in each quarter in 2016 against Zika case, respectively. ARIMAX model was multivariate extension of the ARIMA model and incorporated external effective variable ($Y_t$ in equation 2, representing time series of Zika case counts). The univariate ARIMA model and multivariate ARIMAX model were then compared and assessed whether including external variable actually increased model performance (e.g., decreasing AIC value). Note that ARIMAX model was constructed based on
the corresponding optimal ARIMA model (i.e., ARIMAX and ARIMA model should have exactly the same $p,d,q$ parameter values in order to assess the effect of including external variable). This would reveal whether public engagement to CDC’s tweets (as well as CDC’s own Zika-related tweets) also significantly corresponded to actual Zika epidemic in the U.S. Then we tested whether original CDC tweets, retweets, or replies could serve as imperative indicator of actual Zika case (or vice versa) in different stages by applying Granger causality test. The terms that needed to be first differenced (pre-whitened) in Granger test were determined from the corresponding ARIMA/ARIMAX model (i.e., where parameter $d$ is non-zero).

$$1 - \sum_{i=1}^{p} \phi_i L^i (1 - L)^d X_t = \left(1 + \sum_{i=1}^{q} \theta_i L^i\right) \varepsilon_t + Y_t, \ \text{Eqn. 2}$$

**Results**

**Descriptive and Univariate Time Series Analysis Results**

According to the content analysis, *Zika* (and its related medical condition *microcephaly* in fetus and infant; *Guillain-Barré Syndrome/GBS* in adults) was the 3rd most tweeted health topic by CDC in the entire year 2016, totaling more than 4,000 tweets and was just behind HIV/AIDS and STD (Fig. 1). Among all these Zika-related tweets, a majority of them (more than 60%) were sent out by @CDCgov, the main official Twitter account of CDC, followed by @CDCTravel, @CDCGlobal, and @CDCEmergency. Among all the 67 CDC-associated Twitter accounts (as of 2016), only 21 accounts tweeted about Zika in 2016. Though Zika was one of the key concentrations of CDC, there was substantial temporal heterogeneity in CDC’s tweeting pattern regarding Zika. More than 84% of all Zika-related tweets were sent out in the 1st quarter of 2016, and the subsequent quarters only comprised 5.6%, 7.5%, and 2.4% of all Zika-related
Tweets, respectively (Fig. 2). Zika was unequivocally the most tweeted health topic of CDC in
the 1st quarter and was mentioned in almost 50% of all tweets in that quarter, dwarfing both
HIV/AIDS and STD (and even these two topics combined). This substantial temporal
heterogeneity was also demonstrated by the distinct ARIMA models in each quarter (Table 1, 1st
column for original tweets): the optimal ARIMA model in the 1st quarter was with parameter
$p,d,q=2,0,3$, indicating that the optimal time series model (which minimized AIC value) of
original tweets did not need difference ($d=0$, i.e., already stationary), and with AR and MA term
$p=2$ and $q=3$, respectively. The parameters associated with optimal ARIMA models in the next
three quarters were $p,d,q=2,1,3$ (2nd quarter), $1,1,1$ (3rd quarter), and $2,0,3$ (4th quarter).

Retweets of and replies to the original Zika-related tweets by CDC (shortened as retweets
and replies hereafter) followed overall similar temporal characteristics, where 1st quarter had the
largest number of both retweets and replies. The optimal ARIMA models were again distinct
across four quarters in 2016, for both retweets (Table 1, 2nd column) and replies (Table 1, 3rd
column): the only similarity was retweets in 1st and 2nd quarter, both of which had the same
parameterization ($p,d,q=2,1,3$). Comparing among ARIMA models original tweets, retweets, and
replies, there were only two pairs with the same model parameterization: original and retweets in
2nd quarter (both with $p,d,q=2,1,3$); retweets and replies in 3rd quarter (both with $p,d,q=2,1,2$).
All these results revealed substantial temporal variability across different quarters and among
original tweets, retweets, and replies.

Furthermore, the time series of Zika case in the 50 states of U.S. (domestic case) showed
completely different dynamics than that of original tweets (or retweets, replies). Most notably,
case count in 1st quarter in 2016 was actually much lower comparing to the 2nd and 3rd quarters.
Such discrepancy will be further explored and demonstrated in next section by applying multivariate analyses.

**Multivariate Time Series Analysis Results**

The plots of cross-correlation function between Zika case and each of the following variables: original tweets, retweets, and replies in each quarter of 2016 were provided in Fig. 2-4, respectively. For original tweets and case, strong temporal correlations were observed in 1st, 2nd, and 4th quarter. In 1st quarter, CDC’s tweets regarding Zika preceded actual case for approximately 7-10 days, indicated by the substantial lag of 7, 8, 9, and 10 (Fig. 2 top left panel). In 2nd quarter, CDC’s tweets were ahead of case for approximately two weeks (Fig. 2 top right panel). In 4th quarter, CDC’s tweets were behind Zika case for 1-3 days (Fig. 2 bottom right panel). In 3rd quarters there was no substantial correlation between the two time series. These results revealed that CDC was very active and effective in early warning of Zika epidemic on social media when the actual case number was low, but gradually lost momentum and lagged behind. It also coincided with the results in previous section that CDC’s original Zika-related tweets plunged sharply after 1st quarter in 2016. Similar pattern was observed between retweets and case (Fig. 3), the difference was much stronger temporal correlation in 1st quarter and no substantial correlation in 4th quarter. These results indicated that general public were more engaged in retweeting and helping disseminate the information during the first half of the year. However, replies and case had different pattern (Fig. 4). Replies preceded case counts for about a week in 1st quarter, indicating general public’s willingness in discussing Zika with CDC on Twitter. Then such active engagement also vanished towards later of the year, and by 4th quarter, replies were about 10 days behind actual case.
We also calculated mutual information to explore whether and how we could derive mutual dependence between Zika case and one of original tweets, retweets, and replies, from information perspective (Table 1). In 1st quarter, replies had highest mutual information (0.09) with case, higher than original tweets (0.04) and retweets (0.01). Nevertheless, all these mutual information (Shannon information entropy) were low, indicating potential discrepancy between discussion of Zika on Twitter and actual epidemic. In 2nd quarter, replies, retweets, and original tweets had 0.29, 0.17, and 0.13 mutual information with case, respectively, being the highest mutual information of all four quarters in the year. In 3rd quarter, retweets had largest mutual information with case (0.08), followed by both original and replies tied at 0.02. In 4th quarter, retweet got highest Shannon entropy again (0.07), followed by original tweets and replies with very low mutual information (0.01). These results demonstrated that in general, retweets and replies had more mutual information with Zika case, compared with CDC’s original tweets. Thus, CDC’s tweeting pattern was actually an inferior indicator of Zika epidemic than public engagement to its tweets (retweets and replies).

Mutual information did not consider potential temporal characteristics such as lag or trend, and we further quantified whether including external variable (Zika case counts) could increase the ARIMA model performance (Table 1). In 1st quarter, all ARIMAX models outperformed their ARIMA counterparts by a large margin (dAIC= -2.25, -1.88, and -1.21 for original tweets, retweets, and replies, respectively). Thus, although Zika case counts were lowest in 1st quarter, it still substantially impacted and sparked the temporal dynamics of online discussion of Zika. Then in 2nd quarter, including Zika case only improved ARIMAX model for retweets (dAIC= -0.88); in 3rd quarter, it only improved for replies (dAIC= -0.62); and in 4th quarter, it only improved for original tweets from CDC (dAIC= -0.59). These results also
demonstrated large temporal variability and difference in public engagement versus CDC’s response to Zika.

We further evaluated whether Zika case could be Granger cause of original CDC tweets, retweets, and replies, or vice versa. According to the results from Granger causality test, case count was not Granger cause for original tweets in any quarter, and vice versa. Thus, the correlation between CDC’s tweets regarding Zika and actual case was not strong. Retweets, however, could serve as Granger cause of Zika case for order from 1 up to 5 ($P = 0.05, 0.04, 0.02, 0.01,$ and 0.04, respectively) in 1st quarter. This coincided with previous results that retweets had very high correlation with Zika case in 1st quarter (Fig. 3). Similarly, replies also served as Granger cause in 1st quarter for order 3, 4, and 5 ($P = 0.03, 0.01,$ and <0.001, respectively). Furthermore, replies served as Granger cause again in 4th quarter for order 1 ($P = 0.04$). Interestingly, case counts in 3rd quarter could be Granger cause for replies with order 2 and 3 ($P < 0.001$ for both orders) but not vice versa. This was the only exception that Zika case served as Granger cause for Twitter discussion. Note that Granger causality only provided statistical evidence for potential causality, hence did not guarantee actual causality, i.e., replies as Granger cause in 1st quarter did not mean replies to CDC’s tweets “caused” Zika in U.S.; it should be interpreted that replies preceded Zika case and had strong association with case counts at selected orders. The temporal heterogeneity in Granger test results also showed variability across different quarters.

Discussion

This study is the first of its kind that specifically investigate temporal variability in CDC’s tweeting pattern regarding Zika, and link actual Zika case in the U.S. and general public’s responses through retweeting and replying/commenting. The results demonstrated that there were
substantial discrepancy between CDC’s response to Zika in Twittersphere and Zika epidemic: during the first quarter in 2016 when Zika case counts were low in the U.S., CDC was actually very active in disseminating useful information and was effective in early warning of the upcoming epidemic, including an 1-hour long interactive live chat on Twitter on February 16th 2016. Zika-related tweets comprised almost half of all tweets that CDC sent in the first quarter. Such high level of activity also sparked heated public engagement, as retweets and replies were also synchronized with original tweets, and was the highest in all quarters. However, when Zika case counts started to rise from April 2016, CDC’s tweets related to Zika plunged substantially. Interestingly, while public engagement also decreased dramatically, in 2nd and 3rd quarter of 2016, retweets and replies were actually associated with Zika case as well, as revealed by corresponding ARIMAX models (decreased AIC value compared to ARIMA model, which did not consider Zika case as external input variable). When Florida had high incident case since late July (and even local transmitted case other than travel-related case) and later on summer Olympics in Brazil (August 05 to August 21), there were many retweets of replies, demonstrating public’s increased awareness of this emerging health issue; however CDC’s response to it was not comparable to its tweeting pattern in 1st quarter of 2016.

We have also identified inconsistency among CDC’s 67 official Twitter accounts IDs too, which might have a negative effect for general public to correctly identify legitimate source of health-related information. @CDCgov, @CDC Environment, @CDC_TB, @US_CDCIndia, @CDCHaiti, @DrDeanCDC, @DebHouryCDC, @DrNancy_CDC were all listed as official CDC-associated accounts (https://www.cdc.gov/socialmedia/tools/Twitter.html), yet these IDs were not named consistently and there was not a rule of naming these accounts. And because of such inconsistency, fake or misinformation source might pretend to be a legitimate CDC-
associated Twitter account, for example, we have found @CDCwhistleblowR who claims to be “leading the battle for truth from Big Pharma about vaccines, autism, etc”.

In this study, we focused on dynamics of tweeting/retweeting/replying regarding Zika. Previous studies usually use online social media discussion trend to predict and adjust actual disease dynamics (Nagar 2014, Paul 2014, Adebayo 2017, Bragazzi 2017, McGough 2017) [13,16,18,31]; while it might be useful for locally transmitted diseases such as influenza, we suggest it might not work well for domestic Zika cases in the U.S., as a majority of them were travel-related (highly stochastic) and hence could not be accurately captured by statistical model such as ARIMA or ARIMAX. Nevertheless, the other way around would work (i.e., using actual Zika epidemic to model Twitter discussion dynamics and reveal dynamic change through the year), as we discovered and discussed in this study.

We did not investigate the actual contents of these tweets and replies in this study, and one of the future directions is to explicitly investigate the contents using topic modeling (Miller 2017) [24] and other computational techniques such as natural language processing (Muppalla 2017) [32]. Specifically, because Twitter was still restricted to 140 characters back in 2016, many of the CDC tweets had a link to either an external website or informative figures. We aim to investigate whether including more informative contents in the tweets increase public engagement, comparing to plain text.

While retweeting and replying are considered as different levels of user engagement (Diga 2009) [29], replies reveal more detailed synergistic, neutral, or antagonistic attitude towards CDC’s original tweets. Opinion leaders are able to help disseminate the original information in tweets, or alter general public’s perspective by proposing skepticism, accusation, and even injecting hidden agendas (Miller 2017) [24]. Thus it is imperative to identify opinion
leaders in the information dissemination networks of CDC’s original tweets, and further assess their respective roles in shaping general public’s perceptions. Our preliminary analysis showed that major news outlets (their Twitter accounts, e.g., @CNN, @MSNBC, @FOXNews, @Reuters, etc) and local (state and county) governments (e.g., Florida Department of Health @HealthyFla, California Department of Public Health @CAPublicHealth) were among the top contributors in retweeting and replying to CDC’s tweets. We will also identify other opinion leaders who might work against disseminating useful information to general public, and compare different tweeting behaviors (temporal dynamics in conjunction with content analysis) between federal and local government agencies, news outlets, and the opinion leaders who spread fake or inaccurate information and created confusion, anxiety, and fear in Twittersphere. The insights will help disseminate accurate health- and disease-related information more effectively, and fight against fake health information on social media.

**Limitations**

Certain limitations exist in this study. First, we focused on public engagement (retweeting and replying to) of CDC’s tweets. It was relatively small amount of public engagement comparing to all Zika-related English tweets. Similarly, number of original tweets from CDC were also relatively low especially after 1st quarter, which might influence time series analysis results. A potential remedy was that we showed temporal dynamics of all Zika-related tweets as a comparison in this manuscript. Second, we did not perform content analysis on the tweets/retweets/replies themselves, and it was especially important for replies if we would want to explicitly assess public’s different reactions towards the original tweet (e.g., neutral, synergistic, or antagonistic). This issue will be explored and addressed in future studies.

**Conclusions**
We demonstrated that there was substantial discrepancy among CDC’s tweets regarding Zika, public engagement, and actual Zika epidemic in different stages of epidemic in 2016. CDC sent out more than 84% of Zika-related tweets in 1st quarter of 2016, where Zika case counts were low. Thus CDC was very effective in early warning of Zika in 1st quarter of 2016 and gained substantial public engagement in this time period as well. While Zika case counts increased sharply in 2nd and 3rd quarters, CDC’s efforts on Twitter surprisingly plunged and was not associated with Zika epidemic. Dynamic public engagement was generally different among quarters and was also substantially influenced by and usually preceded Zika epidemic. Public engagement was generally a more prominent predictor of actual Zika epidemic than CDC’s tweets later in the year.

**Abbreviations**

CDC Centers for Disease Control and Prevention

AIC Akaike Information Criteria

CCF Cross-correlation Function

ARIMA Autoregressive Integrated Moving Average

ARIMAX Autoregressive Integrated Moving Average with External Variable

**Acknowledgements:**

We thank Angat Puri and Abhijit Nair, former graduate students at UNC Charlotte who helped collected the Twitter data and performed initial data cleaning. We also thank Data Science Initiative (DSI), UNC Charlotte for archiving and providing the raw Twitter data.

**Authors’ Contributions**

SC performed data analyses, interpreted the results, and wrote the manuscript.

**Conflicts of interest**
The authors declare no conflicts of interest in this study.
References
Table

Table 1. Mutual Shannon Information Entropy, ARIMA/ARIMAX Model Parameters and AIC Values in Different Quarters of 2016

<table>
<thead>
<tr>
<th></th>
<th>Original + Case</th>
<th>RT + Case</th>
<th>Reply + Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1</td>
<td>Mutual Info</td>
<td>0.04</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>ARIMA(X) Par</td>
<td>2, 0, 3</td>
<td>2, 1, 3</td>
</tr>
<tr>
<td></td>
<td>dAIC</td>
<td>-2.25 *</td>
<td>-1.88 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(976.61, 974.36)</td>
<td>(1341.51, 1339.63)</td>
</tr>
<tr>
<td>Q2</td>
<td>Mutual Info</td>
<td>0.13</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>ARIMA(X) Par</td>
<td>2, 1, 3</td>
<td>2, 1, 3</td>
</tr>
<tr>
<td></td>
<td>dAIC</td>
<td>0.96</td>
<td>-0.88 *</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(722.54, 723.50)</td>
<td>(1207.14, 1206.26)</td>
</tr>
<tr>
<td>Q3</td>
<td>Mutual Info</td>
<td>0.02</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>ARIMA(X) Par</td>
<td>1, 1, 1</td>
<td>2, 1, 2</td>
</tr>
<tr>
<td></td>
<td>dAIC</td>
<td>1.95</td>
<td>1.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(719.51, 721.46)</td>
<td>(1172.01, 1173.83)</td>
</tr>
<tr>
<td>Q4</td>
<td>Mutual Info</td>
<td>0.01</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>ARIMA(X) Par</td>
<td>2, 0, 3</td>
<td>0, 1, 2</td>
</tr>
<tr>
<td></td>
<td>dAIC</td>
<td>-0.59 *</td>
<td>1.62</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(453.28, 452.69)</td>
<td>(917.84, 919.46)</td>
</tr>
</tbody>
</table>

Note: negative dAIC value indicates better performance of ARIMAX model comparing to its corresponding ARIMA model (highlighted by *), hence including Zika case counts improve model performance. RT: retweeting without commenting.
Figure Captions and Legends

Top 15 Most Tweeted Health Topics by CDC in 2016

Fig 1. Top 15 most tweeted health topics by CDC in 2016

Note: Zika is #3 resulting in more than 6,000 tweets from all its associated official accounts.

Note there might be overlap between topics (e.g., Zika/STD; Zika/Vaccine; HIV AIDS/PrEP, HPV/Vaccine, etc).
Fig 2. Non-congenital Zika Virus Disease Cases in 50 States/DC and both 50 States/DC and Territories in 2016

Note: data obtained from CDC Morbidity and Mortality Weekly Report (Hall et al. 2018).
Fig 3. Time Series of Zika Related Tweets Sent by CDC, Corresponding Retweets, Replies, and All Original Tweets from CDC in 2016

Note: There is substantial temporal discrepancy between actual Zika case in the U.S. and CDC response in Twitter.
Fig 4. Cross-correlation Function (CCF) between Original CDC Zika Related Tweets and Domestic Zika Case in Four Quarters of 2016
Fig 5. Cross-correlation Function (CCF) between Retweets to CDC Zika Related Tweets and Domestic Zika Case in Four Quarters of 2016
Fig 6. Cross-correlation Function (CCF) between Replies to CDC Zika Related Tweets and Domestic Zika Case in Four Quarters of 2016