


---

## Research and Applications

# Mining Twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the United States

Hansi Zhang<sup>1,†</sup>, Christopher Wheldon<sup>2,†</sup>, Adam G Dunn<sup>3</sup> , Cui Tao<sup>4</sup>, Jinhai Huo<sup>5</sup>, Rui Zhang<sup>6</sup>, Mattia Proserpi<sup>7</sup>, Yi Guo<sup>1</sup>, and Jiang Bian<sup>1</sup>

<sup>1</sup>Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA, <sup>2</sup>Department of Social and Behavioral Sciences, College of Public Health, Temple University, Philadelphia, Pennsylvania, USA, <sup>3</sup>Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia, <sup>4</sup>School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, Texas, USA, <sup>5</sup>Department of Health Services Research, Management and Policy, University of Florida, Gainesville, Florida, USA, <sup>6</sup>Institute for Health Informatics and Department of Pharmaceutical Care & Health Systems, University of Minnesota, Minneapolis, Minnesota, USA, and <sup>7</sup>Department of Epidemiology, University of Florida, Gainesville, Florida, USA

<sup>†</sup>HZ and CW contributed equally to this work.

Corresponding Author: Jiang Bian, PhD, Health Outcomes & Biomedical Informatics, College of Medicine, University of Florida, 2197 Mowry Road, 122, Gainesville, FL 32610, USA; bianjiang@ufl.edu

Received 3 July 2019; Revised 6 September 2019; Editorial Decision 6 October 2019; Accepted 8 October 2019

### ABSTRACT

**Objectives:** The study sought to test the feasibility of using Twitter data to assess determinants of consumers' health behavior toward human papillomavirus (HPV) vaccination informed by the Integrated Behavior Model (IBM).

**Materials and Methods:** We used 3 Twitter datasets spanning from 2014 to 2018. We preprocessed and geocoded the tweets, and then built a rule-based model that classified each tweet into either promotional information or consumers' discussions. We applied topic modeling to discover major themes and subsequently explored the associations between the topics learned from consumers' discussions and the responses of HPV-related questions in the Health Information National Trends Survey (HINTS).

**Results:** We collected 2 846 495 tweets and analyzed 335 681 geocoded tweets. Through topic modeling, we identified 122 high-quality topics. The most discussed consumer topic is "cervical cancer screening"; while in promotional tweets, the most popular topic is to increase awareness of "HPV causes cancer." A total of 87 of the 122 topics are correlated between promotional information and consumers' discussions. Guided by IBM, we examined the alignment between our Twitter findings and the results obtained from HINTS. Thirty-five topics can be mapped to HINTS questions by keywords, 112 topics can be mapped to IBM constructs, and 45 topics have statistically significant correlations with HINTS responses in terms of geographic distributions.

**Conclusions:** Mining Twitter to assess consumers' health behaviors can not only obtain results comparable to surveys, but also yield additional insights via a theory-driven approach. Limitations exist; nevertheless, these encouraging results impel us to develop innovative ways of leveraging social media in the changing health communication landscape.

**Key words:** Twitter, social media, human papillomavirus vaccine, topic modeling, integrated behavior model

---

## INTRODUCTION

Human papillomavirus (HPV) is the most common sexually transmitted disease (STD) in the United States.<sup>1</sup> Although HPV infections are transient, persistent infection can lead to cancer. An estimated 33 700 new patients are diagnosed with HPV-associated cancers (eg, anal, penile, cervical, and oral cancers) each year<sup>2</sup> in the United States. The HPV vaccine is effective in preventing most of these HPV-related cancers for individuals in early age.<sup>3</sup> Nevertheless, in 2017, only 48.6% of U.S. adolescents received recommended HPV vaccination series, and 65.5% received  $\geq 1$  dose of the series.<sup>4</sup> HPV vaccination coverage also varies greatly by state. Only 3 states (ie, District of Columbia: 91.9%, Rhode Island: 88.6%, and Massachusetts: 81.9%) have more than 80% coverage for the first dose, while the bottom 3 states (ie, Kentucky: 49.6%, and Mississippi: 49.6%, and Wyoming: 46.9%) have coverage rates  $< 50\%$ .<sup>4</sup> There is a huge public health need to increase the awareness of HPV-related issues to promote HPV vaccination.

To increase HPV vaccination initiation and coverage, we first need to understand factors that affect people's health behavior toward vaccination uptake. Recognized by the Integrated Behavior Model (IBM), a general theory of behavioral prediction, individuals' intention is the most important determinant of their health behaviors (ie, HPV vaccination uptake in our case), while behavior intention is subsequently determined by attitude (eg, feelings about the behavior), perceived norms (eg, the social pressure one feels to perform the behavior), and personal agency (eg, self-efficacy).<sup>5</sup> Other factors such as knowledge (ie, skills to perform the behavior), environmental constraints (eg, access to care), habits, and salience of the behavior can also directly affect individuals' health behaviors.

Interviews, focus groups, and questionnaires are traditional approaches for understanding these factors that affect individuals' behavior in decision-making processes. A few studies used these traditional approaches to examine the determinants of HPV vaccination uptake.<sup>6–8</sup> The rapid growth of social media has transformed the communication landscape not only for people's daily interactions, but also for health communication. People want their voices to be heard and voluntarily share massive information about their health history and status, perceived value and experience of care, and many other user-generated health data on social media. A few studies also used social media data to understand individuals' HPV vaccination behavior. Du et al<sup>9</sup> leveraged a machine learning–based approach to inspect individuals' attitudes (ie, positive, neutral, and negative sentiments) about different aspects of HPV vaccination (eg, safety and costs) using Twitter data. Keim-Malpass et al<sup>10</sup> mined Twitter data to understand public perception of HPV vaccine through a manual content analysis. Dunn et al<sup>11</sup> explored consumers' information exposure related to the HPV vaccine on Twitter and found that populations disproportionately exposed to negative topics had lower coverage rates. However, very few studies were guided by any well-established health behavior theories.

Shapiro et al<sup>12</sup> used the Health Belief Model to code the types of individuals' concerns such as unnecessary (eg, HPV vaccine is not beneficial), perceived barriers (eg, perceived harms), and cues to action (eg, influential organizations guiding against HPV vaccine), among many other concerns (eg, mistrust, undermining religious principles, undermining civil liberties). However, they did not compare their social media findings with those obtained from traditional methods (eg, surveys). The validity of using social media data for understanding behavioral determinants warrants further investigation.

Further, most of these HPV-related social media studies did not consider the different types of users who posted about HPV: (1) those who are involved in health promotion (eg, government agencies, health organizations, professionals) and (2) individual consumers discussing policies and their own vaccination experiences. While all forms of HPV information may contribute to the factors that shape vaccination behaviors, distinguishing between promotional information and consumers' discussions may help in understanding the impact of health promotion on individuals' behaviors.

In this study, guided by IBM, we aim to assess the determinants (eg, knowledge, attitudes, beliefs) of consumers' HPV vaccination behaviors using Twitter data and compare these Twitter findings with the responses of HPV-related questions in the Health Information National Trends Survey (HINTS).<sup>13</sup> We fill 3 important gaps in prior HPV-related social media studies<sup>14</sup>: (1) we classified HPV-related tweets into promotional information vs consumers' discussions; (2) we mapped the topics learned from consumers' Twitter discussions to IBM constructs; and (3) we assessed the associations between the learned Twitter topics and responses to HPV-related questions in HINTS to determine the feasibility of using social media–derived measures to match or complement survey-based measures of vaccination behaviors. Our study addresses the following research questions (RQs):

1. What are the topics discussed in HPV-related tweets?
2. Are there any correlations between promotional HPV-related information and consumers' discussions on Twitter in terms of topic distributions?
3. Can consumer discussion topics in Twitter be mapped to IBM constructs, and are the geographic distributions of these topics comparable to the determinants measured from HINTS survey?

## MATERIALS AND METHODS

### Data sources

We used 3 Twitter datasets collected independently using Twitter application programming interface with HPV-related keywords. The 3 datasets covered overlapping date ranges, spanning from January 2014 to April 2018 (Table 1). From a total of 2 846 495 tweets, we removed 248 462 duplicates and retained 2 598 033 tweets.

Further, we obtained survey data from HINTS-4-Cycle-4 (ie, covering August 2014 to November 2014) and HINTS-5-Cycle-1 (ie, January 2017 to May 2017). HINTS is a nationally representative survey on public's use of cancer- and health-related information. We extracted responses from 6962 respondents who answered 8 HPV-related questions from the 2 datasets. We obtained state-level geographic information and full-sample weight (ie, to calculate population estimates) of each respondent.

### Data sources

Our data analysis consists of 4 steps (Figure 1) and is detailed subsequently.

#### Step 1: Data preprocessing

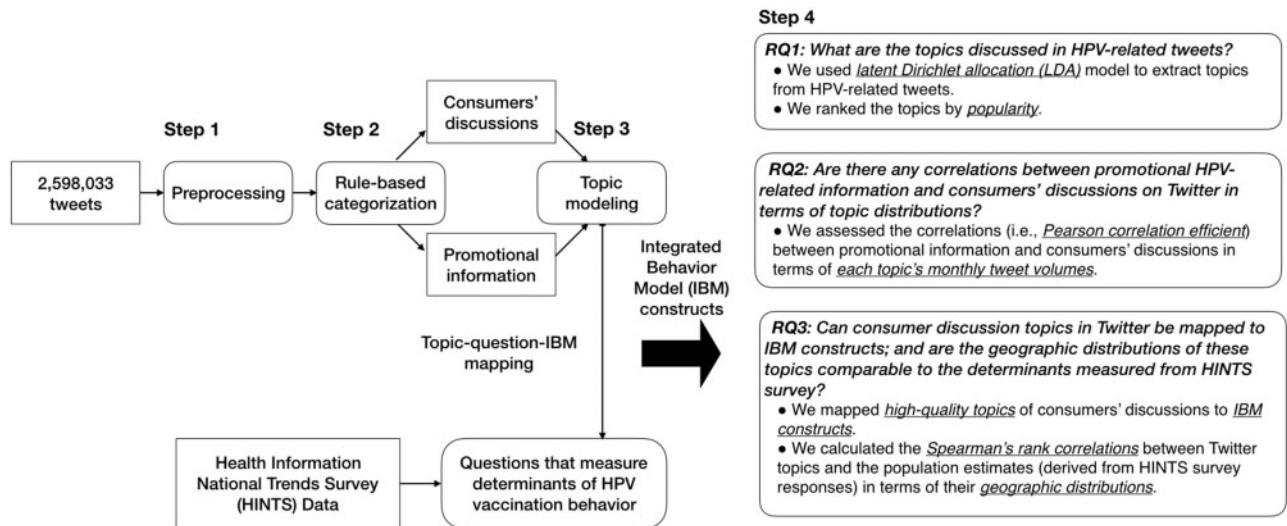
We first removed non-English tweets using a 2-step process. The “*lang*” attribute specifies the language of the tweet, identified by Twitter's internal language detection algorithm.<sup>16</sup> If the “*lang*” attribute was not available, we used Google's language detection algorithm<sup>17</sup> to identify the language. We also made a few other data cleaning

**Table 1.** The 3 HPV-related Twitter datasets, their date ranges, keywords used for data collection, and total number of tweets

Data source	Date range	Keywords <sup>a</sup>	Tweets <sup>b</sup> (N = 2 598 033)
Present study	January 2016 to April 2018	<i>cervarix, gardasil, hpv, human papillomavirus</i>	2 238 433 (86.16)
Dunn <i>et al</i> <sup>11</sup>	January 2014 to December 2016	<i>gardasil, cervarix, hpv + vaccin*, cervical + vaccin*</i>	423 594 (16.30)
Du <i>et al</i> <sup>15</sup>	November 2015 to March 2016	<i>cervarix, gardasil, hpv, human papillomavirus</i>	184 468 (7.10)

<sup>a</sup>“*hpv + vaccin\**” means a tweet has to contain both *hpv* and a word starts with *vaccin*.

<sup>b</sup>Note that there are overlaps across the 3 datasets. The percentage indicates the amount of tweets of each dataset over the total number of unique tweets combined.



**Figure 1.** The overall data analysis workflow. The analysis consists of 4 steps: (1) data preprocessing; (2) rule-based classification of the tweets into either promotional information or consumers' discussions; (3) applying topic modeling to discover major discussion themes and exploring associations between topics in consumers' Twitter discussions and responses to the 8 human papillomavirus (HPV)-related Health Information National Trends Survey (HINTS) questions; and (4) based on these analyses, answering 3 research questions (RQs). IBM: Integrated Behavior Model; LDA: latent Dirichlet allocation.

efforts. We removed (1) hashtag symbols (“#”), (2) uniform resource locators (URLs), and (3) user mentions (e.g. “@username”).

We geocoded each tweet to a U.S. state using a tool we developed previously.<sup>18</sup> Twitter users have 3 options to attach geographic information to their tweets or profiles: (1) a tweet include a geocode (Global Positioning System [GPS] latitude and longitude) or a geographic “place,” if it is posted with a GPS-enabled mobile device or the user chose to tag it with a “place”; (2) the associated user profile can be geocoded (either to a GPS location or a “place”); and (3) the user can fill the “location” attribute with free text.<sup>19</sup> If geocodes are available, we first attempted to resolve the locations through reverse geocoding using GeoNames<sup>20</sup>; otherwise, we matched the free-text “location” with lexical patterns indicating the location of the user such as a state name (eg, “Florida”) or a city name in various possible combinations and formats (eg, “—, fl” or “—, florida”). Very few (ie, 0.85% of all tweets) tweets have GPS geocodes<sup>21</sup>; thus, majority of our tweets were geocoded using lexical patterns.

### Step 2: Rule-based categorization of the tweets

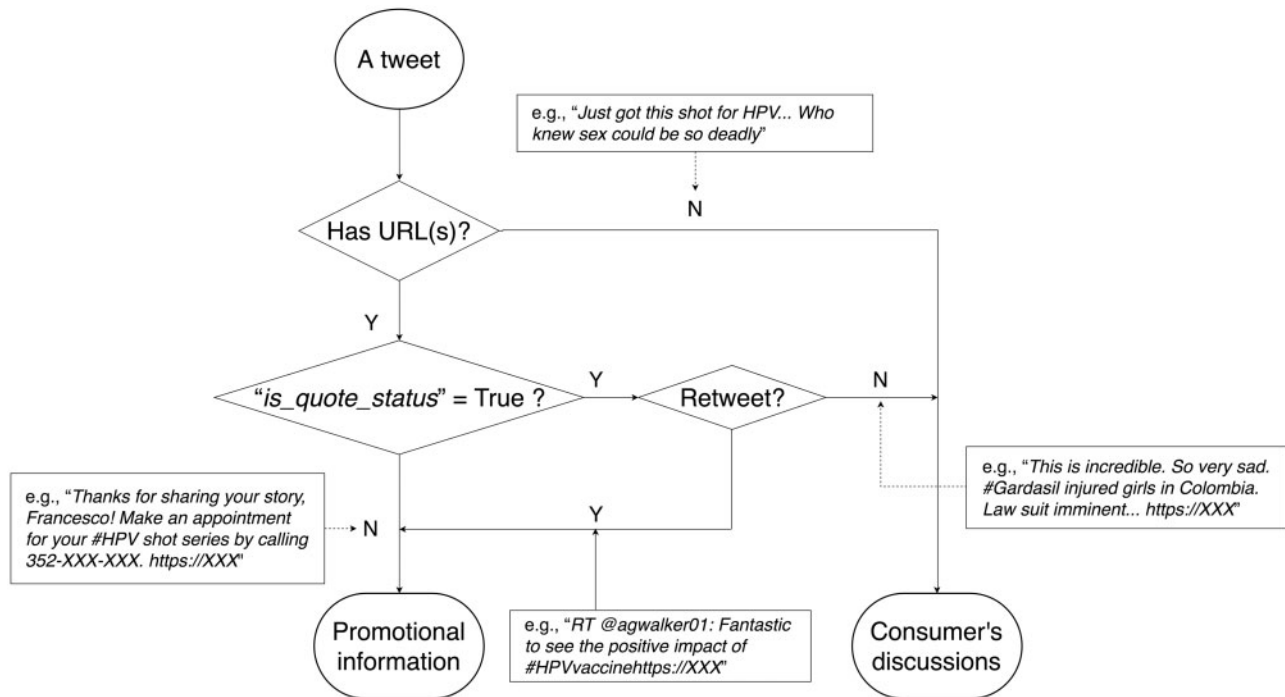
Previously, we built classifiers to filter out irrelevant tweets.<sup>22,23</sup> Nevertheless, in some cases,<sup>24</sup> the keywords used for data collection were specific enough; thus, very few tweets were irrelevant. We randomly annotated 100 tweets and found that only 2 tweets were irrelevant to HPV (ie, 98% were relevant). We thus considered all our tweets as HPV-related without needing a complex classifier.

We then categorized these tweets into either (1) promotional information or (2) consumers' discussions. Consistent with our previous findings,<sup>23</sup> tweets that contain URLs are more likely to be promotional information, in which the URLs are links to HPV-related news, research findings, and health promotion activities. We randomly annotated 100 tweets with URLs and found 95% are promotional information. Further, users can “quote” another tweet or online resources (eg, a web page) but with additional comments expressing their own opinion, and the original quoted tweets (or web pages) are converted into URLs. Twitter users can also “retweet” another tweet (ie, starts with “rt”); nevertheless, the original tweet is not converted into a URL (but URLs in the original tweet were preserved). Based on these observations, we devised a set of simple yet effective rules, as shown in Figure 2. Note that these rules were applied on the original tweets before removing URLs.

### Step 3: Topic modeling

Topic modeling, a statistical natural language processing approach, is widely used for finding abstract underlying topics in a collection of documents. We used the latent Dirichlet allocation (LDA) model<sup>25</sup> to extract topics from our HPV-related tweets. In LDA, each document (ie, a tweet) is modeled as a mixture of topics, and each topic is a probability distribution over words. The LDA algorithm exploits document-level word co-occurrence patterns to discover underlying topics. Based on a prior study, we first removed stop words (eg, *the*, *a*) and words that occurred  $\leq 3$  times in our corpus.<sup>26</sup>

Even though LDA is an unsupervised approach, the number of topics needs to be set a priori. We tested 3 statistical methods to find



**Figure 2.** A rule-based categorization of the tweets into promotional HPV-related information and consumers' discussions. \*If a tweet does not include a Uniform Resource Locator (URL), it is considered as a consumer discussion. Even if it is a retweet (ie, starts with "rt"), the retweet is consumers' discussions, as we considered that the user who retweeted agrees with the original user's discussion and the original tweet is also consumers' discussions (as there is no URL). When a tweet contains URLs, the rules are more complex. First, if a tweet is quoting another tweet or web resource (ie, "is\_quote\_status" = True) and is not a retweet, it is considered as consumers' discussions. In the special case in which the tweet is a retweet of a quoting tweet, we consider this as promotional information because we are unable to determine which of the comments the current user agrees with. In essence, when a tweet is a retweet, we classified the retweet based on the original tweet. Second, if a tweet is not a quoting tweet, it is considered as promotional information. HPV: human papillomavirus.

the best number of topics: (1) Arun et al,<sup>27</sup> (2) Cao et al,<sup>28</sup> and (3) Deveaud et al.<sup>29</sup> However, these methods did not converge on our Twitter corpus. One possible reason is that Twitter messages are short, but the number of tweets is huge; thus, we may need a large number of topics to obtain a reasonable model.<sup>30</sup> Thus, we chose a relatively large number (ie, 150 topics) based on parameters used in similar Twitter LDA studies.<sup>30,31</sup> We also visualized each topic using the top 10 words as a word cloud, in which the size of each word is proportional to its probability in that topic.

The nature of LDA allows all topics (derived from the entire collection of tweets) to occur in the same tweet with different probabilities, while topics with low probabilities might not actually exist. Thus, we needed to determine a cutoff probability value to select the most representative and adequate topics. We tested a range of cutoff values and manually evaluated a random sample of tweets (ie, 100) for each tested cutoff value to determine whether the topics (whose probabilities were larger than the cutoff) assigned to each tweet were correct. We selected the lowest cutoff where more than 80% of topic assignments were adequate. After assigning topics for each tweet, we manually evaluated each topic's word cloud and a sample of associated tweets to determine the topic's (1) theme and (2) quality (ie, a topic was of low quality if more than half of the sample tweets were not relevant to the assigned topic or the word cloud words do not have a consistent theme).

#### Step 4: Research questions

In testing RQ1, to identify popular topics, we calculated percentage of each topic's tweet volume over the total number of tweets for both promotional information and consumers' discussions, and

ranked the topics by popularity (ie, percentage) within each category.

For RQ2, to assess whether promotional information has any impact on consumers' discussions, we assessed the correlations (ie, Pearson correlation efficient) between promotional information and consumers' discussions in terms of each topic's monthly tweet volumes.

For RQ3, we first mapped high-quality topics (step 3) directly to IBM constructs through manually examining each topic's word cloud and a sample of 10 associated consumer tweets (promotional information does not reflect thoughts from lay consumers, thus not considered) by 2 annotators (H.Z. and J.B.). For example, a twitter topic "HPV related cancers" with a sample consumer tweet—"HPV is a contributor to the rise in mouth cancer. . ."—can be mapped to the "knowledge" construct in IBM. A topic is excluded if it does not represent consumers' discussions (ie, more than 5 consumer tweets are irrelevant to the topic theme). Conflicts between the 2 annotators are resolved through discussions with a third reviewer (Y.G.).

We then mapped the high-quality topics to HPV-related HINTS questions. To do so, we first grouped similar HINTS questions into question groups (QGs) and mapped the QGs to IBM constructs. For example, questions "...HPV can cause anal cancer?" and "...HPV can cause oral cancer?" can be grouped into a QG "Knowledge on HPV-cancer relationships" to the "knowledge" construct. We then manually extracted key terms from each survey question and mapped topics to the question based on matching these keywords with the top 20 words in each Twitter topic. For example, *HPV*, *oral*, and *cancer* were extracted from "...HPV can cause oral cancer?" and can be mapped to topic 81 "HPV and oral cancer," where

the top 5 keywords are *oral*, *cancer*, *hpv*, *sex*, and *dentist*. A topic is mapped to a QG if it is mapped to any questions in the group.

To assess whether Twitter data are comparable to survey data in measuring the determinants of vaccination behavior guided by IBM, our first step is to establish the correlations between consumer-related HPV topics and population estimates derived from HPV-related HINTS questions at the state level. To do so, we first aggregated geocoded tweets of the same state and derived the normalized geographic distribution of each topic at the state level (ie, divided the number of tweets for each topic by the total number of consumer tweets in a state). From survey data, to obtain the normalized geographic distribution of HINTS responses, we divided the number of respondents with the answers of interest (eg, responded “Yes” to “. . . HPV can cause anal cancer?” indicating the respondent has the “knowledge”) by the total number of respondents for each state considering each respondent’s full-sample weight<sup>32</sup> in HINTS.

After normalized both Twitter and survey data, we calculated the Spearman rank correlations between Twitter topics and the population estimates (derived from HINTS survey responses to each QG) in terms of their geographic distributions. Note that, considering that we grouped survey questions into QGs, we also combined answers for all questions in that QG (ie, if the respondent responds with the interested answer for any question in that QG).

## RESULTS

### Step 1: Preprocessing

We removed 958 483 non-English tweets and retained 2 598 033, of which 335 681 (12.92%) tweets could be geocoded to a U.S. state for further analysis.

### Step 2: Rule-based categorization of the tweets

We annotated 100 random tweets and assessed the performance of our rules, which achieved a precision of 84.21%, a recall of 86.00%, and a F-measure of 85.10%. We applied these rules on all the geocoded tweets. Of the 335 681 geocoded tweets, 93 693 (27.91%) tweets were classified as consumers’ discussions and 241 988 (72.09%) tweets were promotional information.

### Step 3: Topic modeling

We determined the cutoff probability for topic assignment is .15, where 84% of 100 randomly selected tweets’ topic assignments were adequate. We were able to assign topics to 86.85% (ie, 291 551) of the geocoded tweets. We manually evaluated each topic’s word cloud and 10 random associated tweets to determine its quality, eliminated 28 low-quality topics (of 150), and considered the remaining 122 topics (ie, associated with 281 712 tweets) in further analyses. [Table 2](#) shows example tweets and the topics associated with each tweet. (For all 150 topics summary and corresponding example tweets, see [supplementary appendix](#))

### Step 4: RQs

For RQ1 (What are the topics discussed in HPV-related tweets?), the top 3 topics are visualized as word clouds in [Figure 3](#).

For RQ2 (Are there any correlations between promotional HPV-related information and consumers’ discussions on Twitter in terms of topic distributions?), we plotted the monthly tweet volumes of both promotional information and consumers’ discussions in [Figure 4](#). A total of 87 of the 122 high-quality topics are correlated be-

tween promotional information and consumers’ discussions ( $P < .05$ ) in terms of their monthly volumes. The top 10 correlated topics are presented in [Table 3](#).

For RQ3 (Can consumer discussion topics in Twitter be mapped to IBM constructs; and are the geographic distributions of these topics comparable to the determinants measured from HINTS survey?), we found that 112 of the 122 high-quality topics are relevant to consumers’ discussions and can be mapped to 6 different IBM constructs ([Figure 5](#)): (1) “feelings about behavior” (97 topics), (2) “behavioral beliefs” (92 topics), (3) “normative beliefs—other’s behavior” (36 topics), (4) “knowledge” (23 topics), (5) “normative beliefs—other’s expectation” (7 topics), and (6) “environmental constrains” (2 topics). Note that a topic can be mapped to multiple IBM constructs. The interrater reliability between the 2 annotators is 0.78.

We grouped 8 HPV-related HINTS questions into 5 QGs and mapped the 5 QGs to 3 types of IBM constructs as shown in [Table 4](#). Of the 122 topics, 35 topics were mapped to HINTS questions based on keyword matching through manual review (ie, kappa = 0.72).

We then explored 2 sets of Spearman’s rank correlations between the geographic distributions of (1) the 35 Twitter topics mapped to HINTS QGs based on keyword matching and (2) the 112 topics mapped directly to IBM constructs, and the population estimates derived from HINTS data as shown in [Table 4](#). [Figure 6](#) shows an example of 3 geographic heatmaps for (1) HINTS QG2; (2) topic 17, which was mapped to QG2 by keywords with a low correlation ( $\rho = 0.35$ ,  $P < .05$ ); and (3) topic 127, which was not mapped to QG2 by keywords but had the strongest correlation with QG2 ( $\rho = 0.55$ ,  $P < .01$ ).

## DISCUSSION

In this study, we explored whether user-generated content on Twitter can be used to assess determinants of health behavior, which are traditionally measured through survey questions. We used methods such as topic modeling on HPV-related tweets to answer our 3 RQs.

For RQ1, we found that the most popular HPV-related topics among consumers on Twitter are “cervical cancer screening” and “defunding of planned parenthood,” which account for 24.92% of all consumers’ tweets. The topic “defunding of planned parenthood” is also related to “cervical cancer screening,” as planned parenthood provides 281 063 Papanicolaou tests for cervical cancer screening each year.<sup>33</sup> Further, the popular topics are similar between promotional information and consumer’s discussions, in which 5 of the top 10 topics are the same across the 2.

For RQ2, we found that 87 (of 122) consumer topics are correlated with promotional information, suggesting that promotional health information on Twitter certainly has an impact on consumers’ discussions, which is consistent with our previous study on Lynch syndrome.<sup>23</sup> These strong correlations might, from another perspective, indicate that coordinated national efforts and promotion strategies on raising public awareness of HPV have been rather successful in recent years.

In RQ3, for the 35 topics mapped to HINTS questions by keywords, most of these topics have a negligible correlation (ie,  $\leq 0.3$ ) with HINTS data. One of the 2 highest correlations we found is between topic 55 “HPV vaccine prevents cervical cancer” and QG4: “how successful is HPV vaccine at preventing cervical cancer?” ( $\rho = 0.35$ ,  $P < .01$ ). One potential reason for these low correlations is that the topics learned using LDA can contain multiple themes

**Table 2.** Example tweets and associated topics

Tweet <sup>a</sup>	Top 3 topics <sup>b</sup>
“RT @user1: @ user2 they have had a rise in Anal Cancers due to HPV virus and the fact they think anal sex maintains virginity”	Topic 11, “pap smear test” (19%) Topic 14, “HPV related cancers” (13%) Topic 106, “cervical cancer and death” (11%)
“The startling rise in oral cancer in men - another good reason to vaccinate males against HPV <a href="https://t.co/xxx">https://t.co/xxx</a> ”	Topic 81, “HPV and oral cancer” (20%) Topic 14, “HPV related cancers” (14%) Topic 147, “doctors’ discussions of vaccine” (14%)
“I’m making health calls: HPV infection can cause penile cancer in men; and anal cancer, cancer of the back of the throat.”	Topic 14, “HPV related cancers” (29%) Topic 59, “HPV causes cancer” (11%) Topic 18, “HPV, HPV vaccination and HPV related cancer for man” (8%)
“RT @user1: You don’t have to have sex to get an STD. Skin-to-skin contact is enough to spread HPV. <a href="https://t.co/xxx">https://t.co/xxx</a> ”	Topic 24, “STD and cervical cancer cure” (39%) Topic 56, “questions about HPV and vaccine” (12%) Topic 64, “sex and HPV vaccine” (11%)
“Please join us for a Facebook event about cervical cancer treatments on 1/26 at 2: 00 pm ET <a href="https://t.co/xxx">https://t.co/xxx</a> <a href="https://t.co/xxx">https://t.co/xxx</a> ”	Topic 114, “cervical cancer treatment” (20%) Topic 67, “cervical cancer diagnosis and signs” (11%) Topic 95, “treatment” (9%)
“HPV Vaccine That Helps Prevent Cervical Cancer in Women May Cut Oral Cancer <a href="https://t.co/xxx">https://t.co/xxx</a> ”	Topic 106, “cervical cancer and death” (29%) Topic 0, “blogs about HPV and HPV commercials” (14%) Topic 27, “HPV infection” (10%)
“my doctor accidentally gave me a fourth dose of gardasil so thats where i’m at today”	Topic 147, “doctors’ discussions of vaccine” (18%) Topic 138, “discussions about HPV and vaccine” (15%) Topic 37, “cervical cancer risk” (11%)
“New CDC Recommendations for HPV Vaccines <a href="https://t.co/xxx">https://t.co/xxx</a> ”	Topic 123, “vaccine recommendation” (25%) Topic 85, “HPV vaccine related needs” (14%) Topic 2, “HPV shot / Gardasil shot” (11%)

<sup>a</sup>Tweets are slightly altered to preserve the privacy of the Twitter users without changing the meaning of the original tweets.

<sup>b</sup>Topics and associated probability. Note that the cutoff probability is set to .15; thus, topics with probabilities <.15 were eliminated for each tweet.

(eg, topic 24, “STD and cervical cancer,” contains 2 themes: “STD” and “cervical cancer”). On the other hand, each survey item in HINTS only measures a specific theme (eg, topic 24 was mapped to QG2: “Do you think that HPV is a sexually transmitted disease (STD)?”). Thus, the tweets related to the themes that were not captured in the survey question (eg, “cervical cancer” in this case) are “noises” that lead to a biased correlation measure. To assert the “true” correlations, a method that can further separate each topic into subthemes is needed. Further, depending on what the survey question measures, merely counting the number of tweets in the topic may not yield an accurate measure of the correlation. For example, for survey questions that measure attitude, counting only the tweets that express “attitudes” using sentiment analysis may yield better results.

Furthermore, topics emerged from tweets may provide more insights toward understanding individuals’ attitude and beliefs about HPV vaccination, which are important predictors of their health behavior (ie, HPV vaccination uptake). In Figure 5, topic 14, “HPV-related cancers,” is mapped to the question in QG1 (“Can HPV cause oral cancer?”), in which in its word cloud we found not only words related to “oral cancer” (eg, “throat cancer”), but also keywords related to other cancers (eg, “penile cancer”). Through examining tweets from that topic, we found examples in which users are linking HPV to not only oral cancer, but also other types of cancer (eg, “I’m making health calls: HPV infection can cause penile cancer in men; and anal cancer.”).

For the 112 topics mapped manually to IBM constructs, 45 topics are correlated ( $P < .05$ ) with HINTS responses: 11 with negligible correlations (ie,  $\rho < 0.3$ ), 30 with low correlations (ie,  $0.3 < \rho$

$< 0.5$ ), and 4 topics with moderate correlations (ie,  $0.5 < \rho < 0.7$ ). Most of these topics are related to people’s “attitude” and “perceived norm.” However, constructs such as “personal agency,” “habit,” and “salience of the behavior” are not found in these topics. One possible reason is that compared with “attitude” and “perceived norm,” “personal agency” are more difficult to identify. People may be more willing to talk about their feelings and perceived norms (eg, other’s behavior about HPV vaccination) than they are their own self-efficacy issues in performing the behavior.

In addition, we found that highly correlated topics and HINTS QGs do not necessarily belong to the same IBM constructs. For example, topics that have high correlations with “knowledge” related HINTS questions are all mapped to the construct “attitude.” These are not necessarily “wrong” results. For example, topic 4, “cervical cancer and Andrew’s story” (mapped to “normative beliefs”), is highly correlated with QG4 (mapped to “attitude”). A possible explanation is that the discussion of Andrew’s behavior in fighting cervical cancer can be considered as “normative beliefs—other’s behavior,” which will impact people’s attitude.<sup>34</sup>

### Limitations

First, social media users are different from the general population. Twitter users are younger than the general population.<sup>35</sup> Thus, the representativeness of social media populations should be carefully considered when interpreting study findings. The presence of bots and fake accounts may also distort the representativeness of our findings. Further, the keywords used across the 3 datasets are slightly different, which may lead to data selection bias.

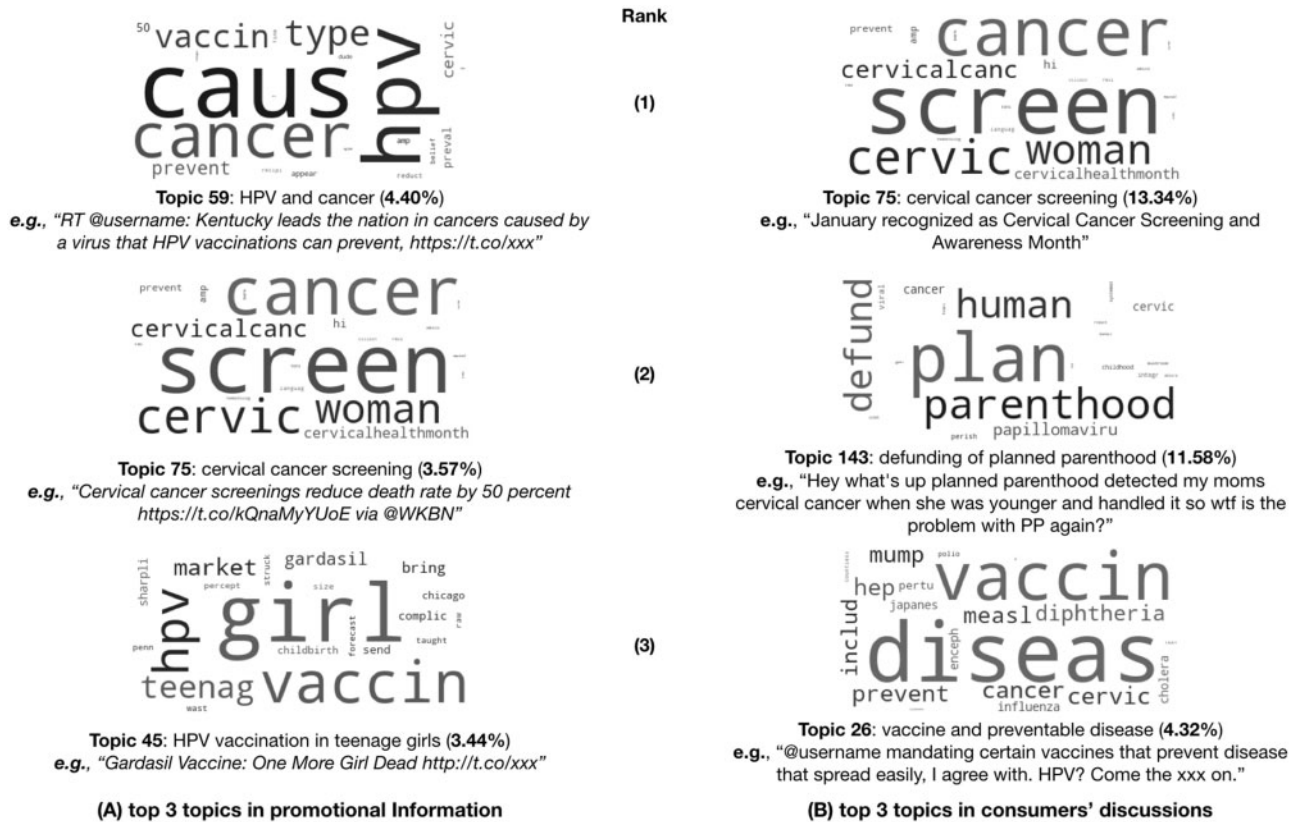


Figure 3. The 3 most popular topics in (A) promotional information and (B) consumers' discussions related to human papillomavirus (HPV) and HPV vaccination.

Table 3. Pearson correlation coefficients between promotional information and consumers' discussions in terms of each topic's monthly tweet volumes

Topic	Correlation coefficient	P value	Tweet volumes <sup>a</sup>	
			Promotional information tweets (n = 241 988)	Consumers' discussions tweets (n = 93 693)
Topic 5, "pap smear test"	0.9517	<.01	5598 (2.31)	2331 (2.49)
Topic 89, "cervical cancer awareness month"	0.9252	<.01	4614 (1.98)	1854 (1.91)
Topic 103, "knowledge of HPV and cervical cancer facts"	0.8758	<.01	5622 (2.32)	1678 (1.79)
Topic 65, "cervical cancer in black women"	0.8096	<.01	5535 (2.29)	2975 (3.18)
Topic 75, "cervical cancer screening"	0.7625	<.01	8628 (3.57)	12 500 (13.34)
Topic 117, "cervical and breast cancer"	0.7608	<.01	2487 (1.03)	1498 (1.60)
Topic 106, "cervical cancer and death"	0.7500	<.01	6896 (2.85)	2772 (2.96)
Topic 14, "HPV-related cancers"	0.7070	<.01	4853 (2.01)	1615 (1.72)
Topic 59, "HPV causes cancer"	0.5247	<.01	10 649 (4.40)	3961 (4.23)
Topic 45, "HPV vaccine in boys and girls"	0.4506	<.01	8320 (3.44)	1915 (2.04)

Values are n (%).

HPV: human papillomavirus.

<sup>a</sup>For clarity, we only presented top 10 correlated topics with tweet volumes >1000. For volume <1000, the sample size might be too small to justify the correlation even with P < .05.

Second, sampling units in social media studies (eg, tweets) are different from traditional survey research (eg, individuals). In Twitter, a user can have multiple relevant posts, and even multiple accounts. Measures derived from counting tweets might be different from surveys that count individual respondents.

Third, the Twitter population and how Twitter users write tweets are constantly changing and evolving. Some of these changes

are triggered by changes made in the Twitter platform. For example, in May 2015, Twitter switched from an opt-out model to an opt-in model for how each tweet can be geotagged, which led to a significant drop of geocoded data in our dataset (ie, from 23.82% to 12.46%, before and after May 2015). Also, in November 2017, Twitter doubled the number of characters from 140 to 280, and we found that 7.03% of our collected English tweets exceeded 140

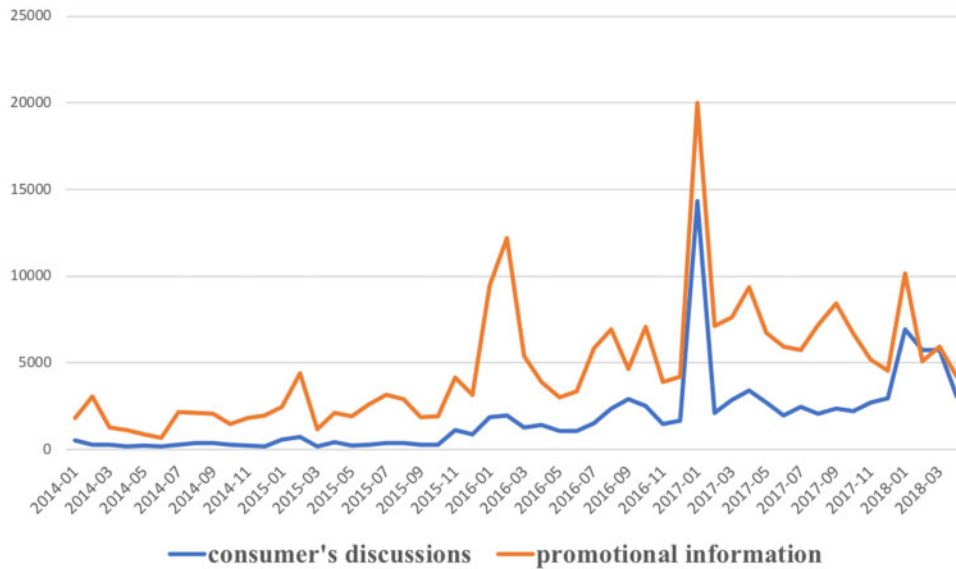


Figure 4. The monthly tweet volumes of promotional human papillomavirus (HPV)-related information and consumers' discussion.

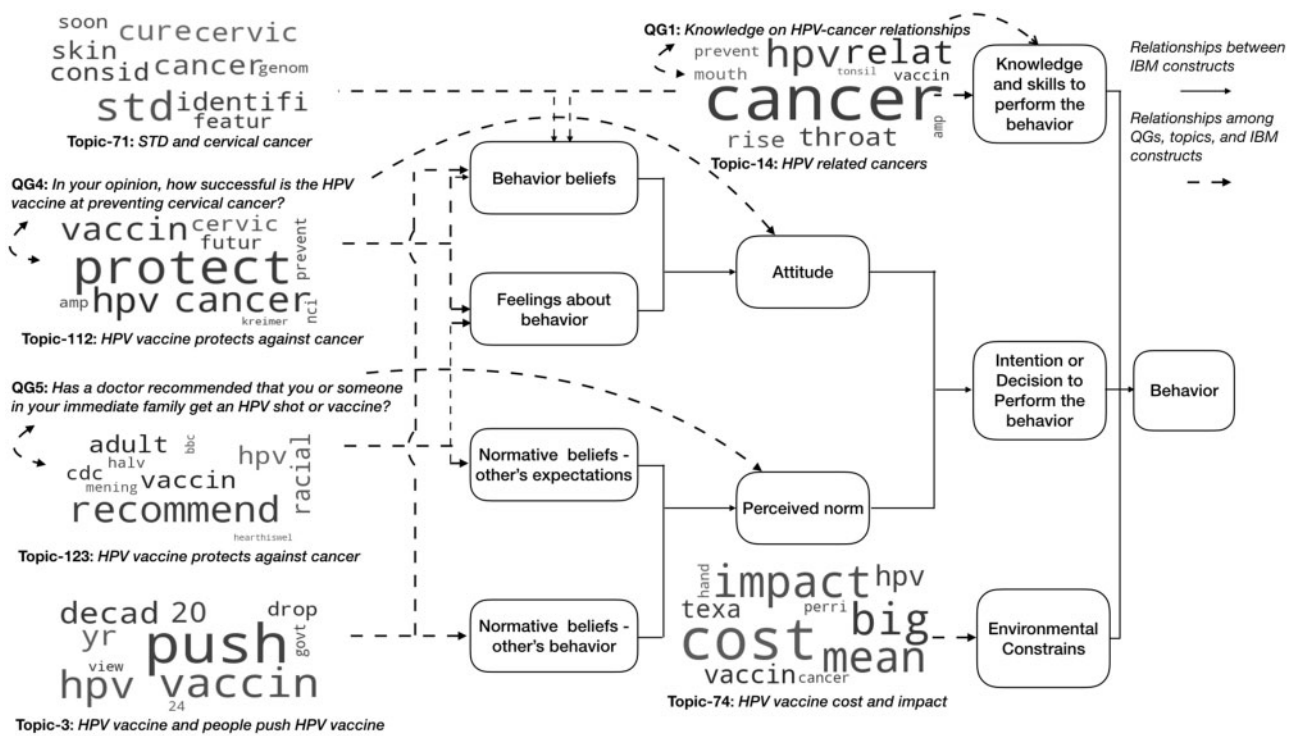


Figure 5. Mapping consumer discussion topics to constructs in the Integrated Behavior Model (IBM), including topics (1) directly mapped to IBM constructs and (2) first mapped to question groups (QGs) and then mapped to IBM constructs (eg, knowledge—QG1, attitude—QG4, perceived norm—QG5). HPV: human papillomavirus.

characters. Further, in June 2019, Twitter announced that they will be removing support for exact geotagging (longitude and latitude), which will have some but limited impact on studies such as ours, as (1) we are looking at the geographic information at the state level and (2) the majority of our tweets (99.15%) are geotagged using lexical patterns based on the “place” or free-text “location” text. Other changes in the Twitter population are natural evolution. For example, according to Pew Research,<sup>36,37</sup> the Twitter population has be-

come even younger than it was previously (ie, Twitter users 18-29 years of age increased from 29% in 2018 to 38% in 2019). Regardless of the reasons for these changes, as we cannot characterize the Twitter population (eg, extract user demographics) with user postings alone, we will not be able to study how these population changes affect our study findings.

Fourth, hashtags are important to help us understand the meaning of a tweet. In this study, we treated hashtags just as another word in the



**Table 4.** Mapping topics in consumers' discussions to the HPV-related survey questions in HINTS and corresponding constructs in the IBM

HPV-related survey questions in HINTS	Mapped topics	IBM construct	Correlation for the 35 topics mapped to HINTS questions by keywords <sup>a</sup>	Correlation for the 112 topics mapped to IBM constructs through manual review (top 3) <sup>b</sup>
QG1. Knowledge on HPV-cancer relationships: a. Do you think HPV can cause anal cancer? b. Do you think HPV can cause oral cancer? c. Do you think HPV can cause penile cancer?	3	Knowledge	Topic 81, "HPV and oral cancer" ( $\rho = 0.29; P < .05$ )	Topic 111, "HPV vaccine and vaccine mandate" ( $\rho = 0.62; P < .01$ ) Topic 127, "HPV symptom and vaccine" ( $\rho = 0.58; P < .01$ ) Topic 124, "vaccine saves lives" ( $\rho = 0.50; P < .01$ )
QG2. Do you think that HPV is a sexually transmitted disease (STD)?	6	Knowledge	Topic 17, "HPV, service and HPV transmission" ( $\rho = 0.35; P < .05$ )	Topic 127, "HPV symptom and vaccine" ( $\rho = 0.55; P < .01$ ) Topic 23, "HPV epidemics" ( $\rho = 0.40; P < .01$ ) Topic 74, "HPV, vaccine cost and impact" ( $\rho = 0.35; P < 0.01$ )
QG3. Do you think HPV requires medical treatment or will it usually go away on its own without treatment?	2	Knowledge	No statistically significant topics	Topic 149, "vaccine victims" ( $\rho = 0.44; P < .01$ ) Topic 9, "fight cervical cancer and hpvax" ( $\rho = 0.35; P < .01$ ) Topic 43, "early detection of cervical cancer" (correlation: 0.27; $P < .01$ )
QG4. In your opinion, how successful is the HPV vaccine at preventing cervical cancer?	26	Attitude	Topic 55, "HPV vaccine prevents cervical cancer" ( $\rho = 0.35, P < .01$ ) Topic 5, "Pap smear test" ( $\rho = 0.32; P < .05$ ) Topic 112, "HPV vaccine protects against cancer" ( $\rho = 0.28; P < .05$ )	Topic 27, "HPV infection" ( $\rho = 0.42; P < .01$ ) Topic 4, "cervical cancer and Andrew's story" ( $\rho = 0.41; P < .01$ ) Topic 146, "HPV, vaccine and sexual behavior" ( $\rho = 0.37; P < .01$ )
QG5. Physician recommendation of HPV vaccination: a. In the last 12 months, has a doctor or health care professional ever talked with you or an immediate family member about the HPV shot or vaccine? b. In the last 12 months, has a doctor or health care professional recommended that you or someone in your immediate family get an HPV shot or vaccine?	2	Perceived norm	No statistically significant topics	Topic 91, "cervical cancer rate and vaccination rate" ( $\rho = 0.44; P < .01$ ) Topic 88, "HPV vaccine and public health" ( $\rho = 0.36; P < .01$ ) Topic 27, "HPV infection" ( $\rho = 0.32; P < .01$ )

IBM: Integrated Behavior Model; HINTS: Health Information National Trends Survey; HPV: human papillomavirus; QG: question group.

<sup>a</sup>Only topics that have significant correlations ( $P < .05$ ) with the survey question groups are listed.

<sup>b</sup>A total of 112 of the 122 high-quality topics were mapped to IBM constructs regardless of whether the topic can be mapped to the survey question group or not.

tweet so that we can capture hashtag words in the topic modeling result. Nevertheless, future studies are warranted to study hashtags separately. Further, as we removed URLs without considering the information posted in the URLs, we might lose some conversational aspects of tweets. For further research, separate and more complex models need to be designed to consider "external content" in those URLs.

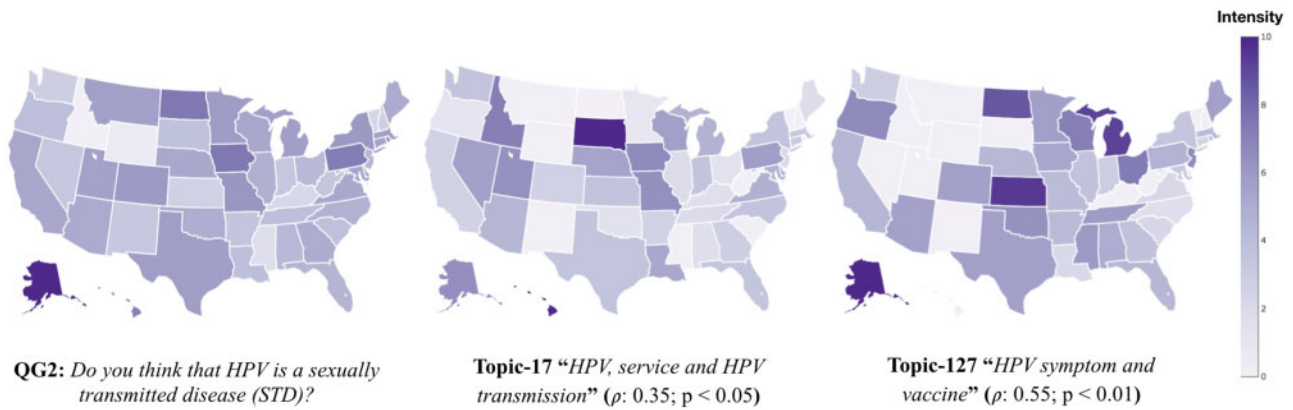
## CONCLUSION

Mining Twitter to assess consumers' health behavior determinants can not only obtain results comparable to surveys, but can yield additional insights via a theory-driven approach. The main contribution of our work is the overall Twitter analysis pipeline and the hybrid use of computational tools (eg, machine learning-based classification and topic modeling approaches) and qualitative methods

(eg, guided by behavioral theory and manual coding of topics). From a consumer health informatics perspective, our approach allows researchers to analyze and understand consumers' needs based on their discussions. Nevertheless, an adequate understanding of the inherent limitations in social media data is always important. Nevertheless, these encouraging results impel us to further develop innovative ways of (1) using social mediadata (eg, to understand factors that are precursors to adopting a health behavioral change) and (2) leveraging social media platforms (eg, to design creative and tailored intervention strategies).

## FUNDING

This work was supported in part by National Institutes of Health grants UL1TR001427 (JB, YG) and R01AT009457 (RZ), the OneFlorida Clinical



**Figure 6.** Geographic heatmaps for the state-level distributions of (1) the responses to Health Information National Trends Survey (HINTS) question group 2 (QG2), (2) the number of tweets in topic 17 that was mapped to QG2 by keywords with a correlation  $\rho = 0.35$  ( $P < .05$ ), and (3) the number of tweets in topic 127 that was NOT mapped to QG2 by keywords but had the strongest correlation with QG2 ( $\rho = 0.55$ ,  $P < .01$ ). The intensity of the color is proportional to the volumes of tweets assigned to that topic or the number of HINTS responses of interest.

Research Consortium (CDRN-1501-26692, JB) funded by the Patient-Centered Outcomes Research Institute, Cancer Prevention & Research Institute of Texas grant 170668 (CT), and National Science Foundation award #1734134 (JB, MP). CW would like to acknowledge the support provided by Dr. Richard Moser and the National Cancer Institute’s Cancer Prevention Fellowship Program. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, Patient-Centered Outcomes Research Institute, or National Science Foundation.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## CONFLICT OF INTEREST STATEMENT

None declared.

## REFERENCES

- Centers for Disease Control and Prevention. Genital HPV Infection-Fact Sheet; 2017. <https://www.cdc.gov/std/hpv/stdfact-hpv.htm>. Accessed August 21, 2018.
- Centers for Disease Control and Prevention. Human Papillomavirus (HPV) Questions and Answers; 2018. <https://www.cdc.gov/hpv/parents/questions-answers.html>. Accessed August 21, 2018.
- Centers for Disease Control and Prevention. HPV Vaccines: Vaccinating Your Preteen or Teen; 2018. <https://www.cdc.gov/hpv/parents/vaccine.html>. Accessed November 28, 2018.
- Walker TY, Elam-Evans LD, Singleton JA, *et al*. National, regional, state, and selected local area vaccination coverage among adolescents aged 13–17 years—United States, 2016. *MMWR Morb Mortal Wkly Rep* 2017; 66 (33): 874–82.
- Glanz K, Rimer BK, Viswanath K, eds. *Health Behavior and Health Education: Theory, Research, and Practice*. 4th ed. San Francisco, CA: Jossey-Bass; 2008.
- Carhart MY, Schminkey DL, Mitchell EM, Keim-Malpass J. Barriers and facilitators to improving Virginia’s HPV vaccination rate: a stakeholder analysis with implications for pediatric nurses. *J Pediatr Nurs* 2018; 42: 1–8.
- Apaydin KZ, Fontenot HB, Shtasel D, *et al*. Facilitators of and barriers to HPV vaccination among sexual and gender minority patients at a Boston community health center. *Vaccine* 2018; 36 (26): 3868–75.
- Sherman SM, Nailor E. Attitudes towards and knowledge about human papillomavirus (HPV) and the HPV vaccination in parents of teenage boys in the UK. *PLoS One* 2018; 13 (4): e0195801.
- Du J, Xu J, Song H-Y, Tao C. Leveraging machine learning-based approaches to assess human papillomavirus vaccination sentiment trends with Twitter data. *BMC Med Inform Decis Mak* 2017; 17 (Suppl 2): 69.
- Keim-Malpass J, Mitchell EM, Sun E, Kennedy C. Using Twitter to understand public perceptions regarding the #HPV vaccine: opportunities for public health nurses to engage in social marketing. *Public Health Nurs* 2017; 34: 316–23.
- Dunn AG, Surian D, Leask J, Dey A, Mandl KD, Coiera E. Mapping information exposure on social media to explain differences in HPV vaccine coverage in the United States. *Vaccine* 2017; 35 (23): 3033–40.
- Shapiro GK, Surian D, Dunn AG, Perry R, Kelaher M. Comparing human papillomavirus vaccine concerns on Twitter: a cross-sectional study of users in Australia, Canada and the UK. *BMJ Open* 2017; 7 (10): e016869.
- Nelson D, Kreps G, Hesse B, *et al*. The Health Information National Trends Survey (HINTS): development, design, and dissemination. *J Health Commun* 2004; 9 (5): 443–60.
- Zhang H, Wheldon C, Tao C, *et al*. How to improve public health via mining social media platforms: a case study of human papillomaviruses (HPV). In: Bian J, Guo Y, He Z, Hu X, eds. *Social Web and Health Research*. Cham, Switzerland: Springer International; 2019: 207–31.
- Du J, Xu J, Song H, Liu X, Tao C. Optimization on machine learning based approaches for sentiment analysis on HPV vaccines related tweets. *J Biomed Semant* 2017; 8 (1): 9.
- Tweet Data Dictionaries. Twitter Developer. <https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object>. Accessed August 21, 2018.
- Shuyo N. Language Detection Library for Java; 2010. <http://code.google.com/p/language-detection/>. Accessed August 21, 2018.
- Jiang B. Twitter User Geocoder. <https://github.com/bianjiang/twitter-user-geocoder>. Accessed August 21, 2018.
- Twitter. Geo Guidelines; 2014. <https://developer.twitter.com/en/developer-terms/geo-guidelines.html>. Accessed December 2, 2018.
- GeoNames. <http://www.geonames.org/>. Accessed October 30, 2017.
- Sloan L, Morgan J, Housley W, *et al*. Knowing the tweeters: deriving sociologically relevant demographics from Twitter. *Sociol Res Online* 2013; 18 (3): 7.
- Hicks A, Hogan WR, Rutherford M, *et al*. Mining Twitter as a first step toward assessing the adequacy of gender identification terms on intake forms. *AMIA Annu Symp Proc* 2015; 2015: 611–20.
- Bian J, Zhao Y, Salloum RG, *et al*. Using social media data to understand the impact of promotional information on Laypeople’s discussions: a case study of Lynch Syndrome. *J Med Internet Res* 2017; 19 (12): e414.
- Bian J, Yoshigoe K, Hicks A, *et al*. Mining Twitter to assess the public perception of the internet of things. *PLoS One* 2016; 11 (7): e0158450.

25. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *J Mach Learn Res* 2003; 3: 993–1022.
26. Schofield A, Magnusson M, Mimno D. Pulling out the stops: rethinking stopword removal for topic models. In: *Proceedings from the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Valencia, Spain: Association for Computational Linguistics; 2017: 432–6.
27. Arun R, Suresh V, Madhavan CEV, Murthy MNM. On finding the natural number of topics with latent Dirichlet allocation: some observations. In: Zaki MJ, Yu JX, Ravindran B, Pudi V, eds. *Advances in Knowledge Discovery and Data Mining*. Berlin, Germany: Springer; 2010: 391–402.
28. Cao J, Xia T, Li J, Zhang Y, Tang S. A density-based method for adaptive LDA model selection. *Neurocomputing* 2009; 72 (7–9): 1775–81.
29. Deveaud R, SanJuan E, Bellot P. Accurate and effective latent concept modeling for ad hoc information retrieval. *Doc Numér* 2014; 17 (1): 61–84.
30. Hong L, Davison BD. Empirical study of topic modeling in Twitter. In: *Proceedings of the First Workshop on Social Media Analytics-SOMA '10*. Washington DC: ACM Press; 2010: 80–8.
31. Sasaki K, Yoshikawa T, Furuhashi T. Online topic model for twitter considering dynamics of user interests and topic trends. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2014: 1977–85.
32. National Cancer Institute. Methodology Reports | HINTS; 2019. <https://hints.cancer.gov/data/methodology-reports.aspx>. Accessed March 23, 2019.
33. Planned Parenthood. Annual Reports. <https://www.plannedparenthood.org/about-us/facts-figures/annual-report>. Accessed January 18, 2019.
34. Ajzen I, Madden TJ. Prediction of goal-directed behavior: Attitudes, intentions, and perceived behavioral control. *J Exp Soc Psychol* 1986; 22 (5): 453–74.
35. Oktay H, Firat A, Ertem Z. Demographic Breakdown of Twitter Users: An analysis based on names. In: *Proceedings of ASE 2014 Big Data Conference at Stanford University*; Accessed September 6, 2018
36. Pew Research Center. Social Media Fact Sheet; 2019. <https://www.pewinternet.org/fact-sheet/social-media/>. Accessed September 5, 2019.
37. Hughes A, Wojcik S. 10 facts about Americans and Twitter. Washington, DC: Pew Research Center; 2019. <https://www.pewresearch.org/fact-tank/2019/08/02/10-facts-about-americans-and-twitter/>. Accessed September 5, 2019.