Geotagged US Tweets as Predictors of County-Level Health Outcomes, 2015–2016

Quynh C. Nguyen, PhD, Matt McCullough, MNR, Hsien-wen Meng, MS, Debjyoti Paul, MS, Dapeng Li, PhD, Suraj Kath, MS, Geoffrey Loomis, MS, Elaine O. Nsoesie, PhD, Ming Wen, PhD, Ken R. Smith, PhD, and Feifei Li, PhD

Objectives. To leverage geotagged Twitter data to create national indicators of the social environment, with small-area indicators of prevalent sentiment and social modeling of health behaviors, and to test associations with county-level health outcomes, while controlling for demographic characteristics.

Methods. We used Twitter’s streaming application programming interface to continuously collect a random 1% subset of publicly available geo-located tweets in the contiguous United States. We collected approximately 80 million geotagged tweets from 603,363 unique Twitter users in a 12-month period (April 2015–March 2016).

Results. Across 3,135 US counties, Twitter indicators of happiness, food, and physical activity were associated with lower premature mortality, obesity, and physical inactivity. Alcohol-use tweets predicted higher alcohol-use–related mortality.

Conclusions. Social media represents a new type of real-time data that may enable public health officials to examine movement of norms, sentiment, and behaviors that may portend emerging issues or outbreaks—thus providing a way to intervene to prevent adverse health events and measure the impact of health interventions. (Am J Public Health. Published online ahead of print September 21, 2017: e1–e7. doi:10.2105/AJPH.2017.303993)

STUDY AIMS AND HYPOTHESES

We created indicators of community sentiment and social modeling of diet, physical activity, and alcohol use. We then tested these sociocultural contextual factors as predictors of county health outcomes. Social learning theory posits that learning is a cognitive process that occurs in a social context. Views and activities described via social media can help shape perceived norms, attitudes, beliefs, and, subsequently, behaviors of people.

We hypothesized that communities that are happier, more actively model healthy eating and physical activity, and have lower...
social displays of alcohol-use behaviors will have lower mortality, obesity, and alcohol-use problems as well as higher levels of physical activity. We used social media data as a means to tap into the attitudes, norms, and behavioral control activities of a community. In an effort to share the results of our data analysis, we created an interactive Web-based mapping application by using open-source technology that allows the public to explore aggregated data by county.

**METHODS**

We used Twitter’s streaming application programming interface to collect a random 1% subset of publicly available, geotagged tweets across the contiguous United States (excluding AK and HI) between April 2015 and March 2016. We dropped duplicate tweets according to their “tweet_id” (each tweet has a unique identification). We removed job postings according to the hashtags “#job” and “#hiring.” We manually examined outliers in our data sets (the top 99th percentile of tweeters) and eliminated automated accounts and accounts for which the majority of tweets were advertisements. Postprocessing resulted in the removal of approximately 1% of tweets with a final analytic data set comprising 79,848,992 geotagged tweets from 603,363 unique users. Geotagged tweets have latitude and longitude coordinates, which enable spatial mapping to their respective county locations. To accomplish spatial join of the tweets, we utilized Python programming language (Python Software Foundation, Wilmington, DE) and the Python R-tree library to build a spatial index. More details of our methodology can be found in Nguyen et al.

**Sentiment Analysis**

We used a Java-based package for natural language processing, MAME Learning for Language Toolkit (MALLET; McCallum, Amherst, MA), for sentiment analysis. We used labeled tweets to train the maximum entropy text classifier in MALLET. MALLET estimates predicted probabilities that a tweet is happy according to word-level features. The classifier uses search-based optimization to assign weights that maximize the likelihood of the labeled training data. However, unlike naïve Bayes, the maximum entropy classifier does not assume conditional independence among features. To train the classifier, we obtained training sets from Sentiment140 (n = 1.6 million tweets), Sanders Analytics (n = 55,133 tweets), and Kaggle (n = 70,866 tweets).

MALLET assigned to each tweet a predicted probability from 0 to 1.0 that the tweet was happy. During our pilot testing of the classifier, we manually labeled a random subset of 1200 tweets as “happy” or “not happy.”

Increasing the MALLET score improved the accuracy against human annotations, but also reduced the calculated prevalence of tweets deemed as “happy.” A MALLET cutpoint of 0.80 achieved the highest level of accuracy while still maintaining a prevalence of happy tweets of 19% (which approximates the prevalence obtained by human annotators).

**Food and Physical Activity Tweets**

We created a list of 1430 popular foods to track the frequency of their social media mentions. Each food item was associated with a measure of caloric density, operationalized as calories per 100 grams based upon data from the US Department of Agriculture national nutrient database. We labeled fruits, vegetables, nuts, and lean proteins (e.g., fish, chicken, and turkey) as “healthy foods.” We excluded fried foods from our count of healthy foods. We also tracked alcohol mentions by using 66 terms that included popular alcoholic beverages (e.g., martini) and alcohol types (e.g., wine, beer, and liquor). Our algorithm excluded phrases that contain alcohol-related terms but refer to nonsubstance objects (e.g., margherita pizza, root beer).

To track physical activity tweets, we created a list of 376 physical activities gathered from physical activity questionnaires, compendia of physical activities, and popularly available fitness programs. Our physical activity list comprised 376 different activities that incorporate gym-related exercise (e.g., treadmill), sports (e.g., baseball), recreation (e.g., hiking), and household chores (e.g., gardening). We excluded popular phrases that generally do not relate to physical activity such as “walk away.” For team sports, we required that the tweet include the words “play,” “playing,” or “played,” which further enabled differentiation between playing a sport and watching a game. Physical activity tweets comprised a mixture of tweets that were about intentions, desire, and reporting on current and past engagement (e.g., tweets about being at the gym or having gone to the gym).

For quality control, two authors (Q. C. N. and H. M.) manually labeled 5000 food and physical activity tweets. These tweets were distributed as follows: food-related (2000), non–food-related (500), physical activity–related (2000), or non–physical-activity–related (500). Among the algorithm labeled food-related tweets, 83% were labeled accurately when compared with labels generated by manual categorization. Similarly, among the algorithm labeled non–food-related tweets, 81% were labeled accurately. Overall, accuracy for food tweets was 83% and the F-score was 0.86. In addition, of the algorithm labeled physical-activity–related tweets, 82% were labeled accurately when compared with labels generated by human categorizers. The accuracy of the algorithm in labeling non–physical-activity–related tweets was 97%. The F-score was 0.90 and the overall accuracy was 85% for physical activity tweets. Typical errors in classification included the use of a figure of speech (e.g., running late, sweet as honey) or a reference to watching sports games rather than playing sports.

We further evaluated our sentiment analysis activities through Amazon Mechanical Turk. We randomly selected 500 tweets (50% labeled as happy and 50% as not happy by our algorithm). Then we created 20 online surveys through random sorting, with each survey consisting of 25 tweets. Participants rated the sentiment of each tweet. Surveys were live from April 1, 2015, to April 5, 2015, and automatically closed with 15 responses. Each tweet was assigned a label (“happy” or “not happy”) on the basis of the modal responses. The accuracy for labeling of happy tweets and nonhappy tweets was 69% and 80%, respectively. The overall sentiment accuracy was 78%, with an F-score of 0.54.

**County-Level Health Outcomes**

We then aggregated all Twitter-derived data to the county level to compare with county-level health outcomes. We obtained county health data from external sources that age-adjusted measures to the 2000 US

---

**e2 Research and Practice**  Peer Reviewed  Nguyen et al.  AJPH  Published online ahead of print September 21, 2017
standard population. Data for premature mortality came from the Centers for Disease Control and Prevention WONDER (Wide-Ranging Online Data for Epidemiologic Research) mortality data (2011–2013). We defined premature mortality per 100,000 as deaths occurring before age 75 years. We obtained data on alcohol-impaired driving deaths for the years 2010 to 2014 from the Fatality Analysis Reporting System. Alcohol-impaired driving deaths was the percentage of motor vehicle crash deaths with alcohol involvement.

We obtained data on chronic conditions and health behaviors from the 2011 to 2014 Behavioral Risk Factor Surveillance System. We assessed adult obesity by the percentage of the adult population (aged 20 years and older) that reported a body mass index of 30 kilograms per meter squared or more. We assessed physical inactivity by the percentage of adults aged 20 years and older reporting no leisure-time physical activity in the past month. We defined excessive drinking as the percentage of adults reporting heavy drinking (drinking more than 1 [women] or 2 [men] drinks per day on average) or binge drinking (consuming more than 4 [women] or 5 [men] alcoholic beverages on a single occasion in the past 30 days).

### Analytic Approach

We categorized Twitter characteristics into tertiles—high, moderate, and low (referent category). In adjusted linear regression models, we used Twitter-derived indicators to predict health outcomes across more than 3135 US counties, while controlling for demographic characteristics. We ran models separately for each health outcome. Sample size varied because of missing outcome or predictor variables. The median number of tweets for county estimates was 2530. We obtained county-level demographic characteristics from the 2010 to 2014 American Community Survey 5-year estimates and they included the following: median age, percentage non-Hispanic White, and median household income to capture information on compositional and economic characteristics of a community.

We evaluated statistical significance at \( P < .05 \). To account for spatial autocorrelation, we adjusted standard errors for clustering of county values within a state. We performed data processing and statistical analysis tasks with Stata MP13 (StataCorp LP, College Station, TX).

We built an interactive Web-based mapping application to visually display study data at the county level. The online mapping application was built by using custom hypertext markup language, cascading style sheets, CARTO cloud software (CARTO, New York, NY) and Google Maps JavaScript application programming interface. The custom mapping application, county-level Twitter data set, and data dictionary are hosted on Github: https://hashtaghealth.github.io/geoportal/start.html.23

### RESULTS

Table 1 shows descriptive statistics for Twitter-derived characteristics aggregated to the county level. Across 3135 US counties, the average prevalence of happy tweets was about 19%. On average, 4% of tweets mentioned food (Table 1). Among these food tweets, the average caloric density of the mentioned food was approximately 240

<table>
<thead>
<tr>
<th>County Level Characteristics</th>
<th>No. of Tweets</th>
<th>No. of Counties</th>
<th>Mean ± SD</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Happiness</strong>: % of tweets that are happy</td>
<td>79 848 992</td>
<td>3 135</td>
<td>18.54 ± 6.29</td>
</tr>
<tr>
<td><strong>Food culture</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Calories density of food tweets (cal/100 g)</td>
<td>4 041 521</td>
<td>3 058</td>
<td>238.23 ± 65.12</td>
</tr>
<tr>
<td>% of tweets about food</td>
<td>4 041 521</td>
<td>3 058</td>
<td>3.85 ± 2.40</td>
</tr>
<tr>
<td>% of tweets about healthy foods</td>
<td>4 041 521</td>
<td>2 900</td>
<td>0.78 ± 0.84</td>
</tr>
<tr>
<td>% of tweets about fast food</td>
<td>4 041 521</td>
<td>2 387</td>
<td>0.33 ± 0.27</td>
</tr>
<tr>
<td>Sentiment of food tweets, % happy</td>
<td>4 041 521</td>
<td>3 058</td>
<td>25.03 ± 12.00</td>
</tr>
<tr>
<td>Sentiment of healthy food tweets, % happy</td>
<td>644 489</td>
<td>2 900</td>
<td>24.64 ± 18.25</td>
</tr>
<tr>
<td>Sentiment of fast-food tweets, % happy</td>
<td>373 449</td>
<td>2 387</td>
<td>16.56 ± 19.09</td>
</tr>
<tr>
<td><strong>Physical activity culture</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% of tweets about physical activity</td>
<td>1 473 984</td>
<td>3 055</td>
<td>2.08 ± 2.09</td>
</tr>
<tr>
<td>Sentiment of physical activity tweets, % happy</td>
<td>1 473 976</td>
<td>3 055</td>
<td>25.63 ± 14.13</td>
</tr>
<tr>
<td><strong>Substance use</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>% tweets about alcohol</td>
<td>687 496</td>
<td>2 769</td>
<td>0.68 ± 0.77</td>
</tr>
<tr>
<td>% tweets about drugs</td>
<td>687 496</td>
<td>1 779</td>
<td>0.08 ± 0.10</td>
</tr>
<tr>
<td>% tweets about smoking</td>
<td>687 496</td>
<td>998</td>
<td>0.07 ± 1.06</td>
</tr>
<tr>
<td>Sentiment of alcohol tweets, % happy</td>
<td>638 347</td>
<td>2 770</td>
<td>27.65 ± 21.03</td>
</tr>
<tr>
<td><strong>County health outcomes</strong>[^3]</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Premature mortality[^4], per 100 000</td>
<td>\ldots</td>
<td>2 989</td>
<td>8 025.59 ± 2 409.21</td>
</tr>
<tr>
<td>% obesity</td>
<td>\ldots</td>
<td>3 142</td>
<td>30.73 ± 4.41</td>
</tr>
<tr>
<td>% diabetes</td>
<td>\ldots</td>
<td>3 220</td>
<td>9.70 ± 2.19</td>
</tr>
<tr>
<td>% leisure-time physical inactivity</td>
<td>\ldots</td>
<td>3 142</td>
<td>25.58 ± 4.93</td>
</tr>
<tr>
<td>% binge or heavy drinking[^5]</td>
<td>\ldots</td>
<td>3 140</td>
<td>16.63 ± 3.36</td>
</tr>
<tr>
<td>% driving deaths with alcohol involvement</td>
<td>\ldots</td>
<td>3 118</td>
<td>31.36 ± 15.91</td>
</tr>
</tbody>
</table>

[^3]: Twitter data collection period: April 2015–March 2016. County summaries of happiness were derived from 80 million tweets from the contiguous United States. Food indicators were derived from 4 million food tweets. Physical activity indicators were derived from 1.5 million physical activity tweets. Substance use indicators were derived from about 700,000 substance-related tweets.


[^5]: We defined premature mortality as deaths occurring before age 75 years.

[^6]: Heavy drinking was defined as drinking more than 1 (women) or 2 (men) drinks per day on average; binge drinking as consuming more than 4 (women) or 5 (men) alcoholic beverages on a single occasion in the past 30 days.
calories per 100 grams. Tweets about healthy foods were happier than those about fast food (25% vs 17%). On average, 2% of tweets were about physical activity and less than 1% of tweets mentioned alcohol use. Tweets about alcohol use were slightly happier than those about physical activity or healthy foods (Table 1). Prevalence of happy tweets correlated with prevalence of food tweets ($r = 0.39$), physical activity tweets ($r = 0.29$), and alcohol tweets ($r = 0.23$). Spatial autocorrelation analyses found that Moran’s $I$ was 0.23 for Twitter happiness, 0.18 for food, 0.11 for alcohol, and 0.14 for physical activity tweets.

Table 1 also shows descriptive statistics on county health outcomes. At the county level, the average premature mortality rate was approximately 8000 per 100 000. The average obesity rate among all counties was 31% and the average percentage of physically inactive adults was slightly more than 25%. The average percentage engaging in excessive drinking neared 17% and close to one third of driving deaths involved alcohol at the county level (Table 1).

Analyses examining temporal trends (Figures A to F, available as supplements to the online version of this article at http://www.ajph.org) found that healthy food mentions were highest in June, July, August, and November, which reflects the possible greater abundance of available fresh fruits and vegetables during those times of the year. Caloric density of food tweets was highest in November, December, January, February, and April—months with major national holidays. Caloric expenditure related to physical activity tweets were lower in the spring and summer months. The lowest prevalence of happy tweets (15%) occurred in April.

Figure 1 presents the spatial distribution of food tweets across the 48 contiguous states and the District of Columbia. Our data suggest that the lowest prevalence of food tweets was in southern states (MS, AL, LA, and OK), West Virginia, and North Dakota. For physical activity tweets, Montana, Arizona, Wyoming, Utah, and Maine had the highest prevalence of physical activity mentions (Figure G, available as a supplement to the online version of this article at http://www.ajph.org). Happy tweets were least frequent in Louisiana, North Dakota, Oregon, Maryland, Texas, Delaware, West Virginia, and Ohio.

Table 2 and Figure 2 display the results of adjusted linear regression analyses examining associations between Twitter-derived county characteristics and county-level health outcomes. Across the range of health outcomes, $R^2$ varied from 0.33 to 0.47. County-level median age was positively associated with lower premature mortality whereas percentage non-Hispanic White and median household income were negatively associated, but only median household income was statistically significant. Greater percentages of happy, food, and physical activity tweets were negatively associated
with lower premature mortality. For instance, counties with the most tweets about physical activity experienced 714 fewer premature deaths per 100,000 compared with counties with the fewest physical activity tweets (Figure 1).

Moreover, counties with the greatest percentages of happy, food, and physical activity tweets had obesity prevalences that were lower by 2.23% to 2.49% (Table 2).

Counties with the highest happy, food, and physical activity tweets had physical inactivity prevalences that were 1.97% to 3.62% lower. Higher percentages of healthy food tweets and lower-calorie food tweets were also statistically significantly associated with better health outcomes (not shown). In addition, positive sentiment toward healthy foods and physical activity were found to be significantly associated with lower obesity and physical inactivity (Table A, available as a supplement to the online version of this article at http://www.ajph.org). We also examined Twitter mentions of alcohol use and its relationship to county-level alcohol-related outcomes (Table B). Counties with the highest tertile of alcohol-related tweets had 3.65% more driving deaths with alcohol involvement and 2.26% more of the population engaging in binge drinking or heavy drinking, compared with counties with the fewest alcohol-related tweets (Table B).

**DISCUSSION**

We used 80 million geotagged tweets from publicly available Twitter data over a 1-year period to create indicators of the social environment. There were 3 major findings in the study. First, social modeling of behaviors on Twitter around food and physical activity were associated with lower mortality, obesity, and physical inactivity at the county level. Second, happiness and positive sentiment around healthy behaviors were linked to better health outcomes. Third, greater prominence of a culture of substance use as indicated by higher social media mentions of alcohol use was related to higher rates of excessive drinking and alcohol-related mortality. Twitter characteristics were predictive of health outcomes, with control for demographic and economic composition.

**Study Findings in Context**

People use Twitter to share news, opinions, and information about their activities. For instance, people often tweet about the food they are about to eat, which reflects their dietary choices and is linked to obesity and diabetes risk. Twitter indicators of happiness and better health behaviors indicated worse patterns for the South—which is affected by higher poverty, less access to resources including health care, and worse health outcomes.

In addition, our study highlights the potential influence of social processes. Norms, values, assumptions, and health beliefs enriched in the social environment can have an impact on the development and maintenance of behaviors. Our finding that...
alcohol mentions on social media were associated with county alcohol behaviors and mortality is aligned with research documenting the influence of social networks.26 Rosenquist et al.’s analysis of a large social network found that changes in the alcohol consumption of a person’s social network was associated with changes in that person’s subsequent alcohol-related behavior.27 In this study, we found that counties with the lowest alcohol tweets had 3.65% fewer alcohol-related deaths. In 2014, 9,967 people died in alcohol-impaired driving deaths. A reduction of 3.65% fewer alcohol-related driving deaths would translate into about 113 fewer deaths.28

Research into the health effects of happiness is nascent; however, there are ties between happiness and mortality—with some indication that the links may be partially mediated by happiness’s influence on health behaviors.29 Psychosomatic theories posit that people consume comfort foods or engage in emotional eating to combat symptoms of psychological distress. Similarly, negative emotions have been found to be associated with irregular physical inactivity—possibly by lowering motivation.30 In our study, we found that happy tweets and physical activity tweets were associated with a 2% to 3% reduction in county-level physical inactivity. Previous research has found that engaging in even low levels of physical activity as compared with being inactive is related to substantial declines (20%) in mortality.31

Study Strengths and Limitations

In the study, we used innovative methods to capture social–environmental features with potential impacts on health. Twitter-derived characteristics were related to important population-based measures of morbidity and mortality. Twitter and other social media platforms offer a way for health researchers to gauge the “pulse” of a community by analyzing online expressions, opinions, and sharing of information.

One strength of using geotagged social media data is that the latitude and longitude coordinates of tweets can be aggregated to other boundaries including census tracts, zip codes, and neighborhood definitions from local planning agencies. In this analysis, we examined associations at the county level. Analyses at a different level may lead to different associations or strength of associations (modifiable area unit problem). For instance, in-progress work with state health outcomes suggests stronger associations between Twitter-derived characteristics and state mortality and chronic conditions than seen at the county level. Analyses at smaller levels of geographies such as census tract and zip code were hindered by the lack of publicly available national data on health outcomes at small geographies.

Construction of neighborhood indicators required that we restrict our data collection to geotagged tweets—tweets in which users enabled location on their mobile phones. Previous studies suggest that about 1% to 2% of tweets may contain GPS location information.32 Users who enable geotagging of their tweets differ demographically from those who do not; for instance, they are slightly older and more likely to be male, but these differences are small.33 Another limitation is the non-representativeness of Twitter users to the general US population. Only 23% of all Internet users and 20% of the US adult population use Twitter.34 Twitter is more popular among online individuals living in urban areas than in rural areas (30% vs 15%), and among online adults younger than 50 years versus those aged 50 years and older (30% vs 11%). Although Twitter data may not be representative of the general population, nonetheless, online expressions of users may have utility in providing information on shared environmental features of the community at large.

For the sentiment analysis, the model was only able to process English-language tweets, thus possibly limiting conclusions to English speakers. Cultures differ with regard to their happiness and verbal expression of happiness, with, for example, some cultures having norms that more strongly encourage expressing positive emotion.35 Cultures emphasizing individuality express emotions differently from those that emphasize group harmony.36 Our sentiment analysis targeted sentiment classification as “happy” versus “not happy” (encompassing both neutral and sad emotions). Thus, we were not able to specifically examine the prevalence of sad tweets. In future work, we plan to identify negative-affect expressions on social media and examine their relationship to health outcomes. Despite these limitations, social media represents a cost-efficient data resource for the construction of contextual features that may have bearing on health outcomes. Future directions may further explore the potential of delivering public health interventions through social media and of further utilizing social media for community health needs assessment.

Public Health Implications

Our analysis indicated that county-level Twitter characteristics were linked with important health indicators such as premature mortality, obesity, and health behaviors. Emerging sources of big data such as social media offer new opportunities for measuring and assessing the public health needs of different communities. These new types of real-time data sources may enable public health officials to examine movement of norms, sentiment, and behaviors that may portend emerging issues or outbreaks—thus providing officials a way to intervene to prevent adverse health events and also to measure the impact of health interventions.

CONTRIBUTORS

Q. C. Nguyen took the lead in designing the study, implementing analyses, and writing the article. M. McCullough mapped the data, created the online mapping tool, and edited the article. H. Meng assisted with the analyses and edited the article. D. Paul managed the Twitter database and edited the article. D. Li performed the spatial joins and geocoding and edited the article. S. Rath collected the tweets and implemented the computer algorithms to produce indicators of sentiment, diet, and physical activity. G. Loomis assisted with the concept of the article and editing the article. E. O. Neessie, M. Wenz, K. R. Smith, and F. Li assisted with the design of the study and edited the article.

ACKNOWLEDGMENTS

This study was supported by the National Institutes of Health’s Big Data to Knowledge Initiative grants 5K01ES025433 and 3K01ES025433-03S1 (PI: Q. C. N.). We thank James VanDerslice, PhD, for his guidance on the project. We thank Minh Pham for her assistance with the online geoportal.

Note. The funder did not have any role in the study design; in the collection, analysis, and interpretation of data; in the writing of the report; and in the decision to submit the article for publication.

HUMAN PARTICIPANT PROTECTION

The University of Utah institutional review board approved the study.

REFERENCES
