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 3135 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)

Geography is an important determinant of health. Where we live, including the social, political, economic, and built environment, has an impact on health and creates health inequities.[1–3] County-level contextual factors (e.g., socioeconomic status, public health policy, and access to health care) have been associated with coronary heart disease,[4] health-related quality of life,[5] and obesity.[6] The built environment plays an important role at the county level. For example, density of fast-food restaurants has been associated with higher individual-level weight.[7] Social processes and networks can also affect health through mechanisms such as the maintenance of norms around health behaviors and the stimulation of new interests. However, patterns observed in one area may not be applicable to another as characteristics vary by location. One way to understand disparities is through the use of pervasive and publicly available social media data.

The widespread use of the Internet and openly shared personal opinions with geotagged check-ins enable researchers to understand real-time local area interactions and perform public health surveillance activities. Twitter is one of the most popular social media platforms in use today. Because of the widespread use of social media, the data can be effectively used to discover patterns and emerging health-related issues. Nascent research has suggested that Web searches and social media can provide up-to-date detection, tracking, and predictions of disease outbreaks.[8] Twitter has been used by researchers and public health agencies to track foodborne illness[9] and for real-time detection of natural disasters and disaster response.[10]

Information generated via Twitter can be useful in the examination of various health-related issues, such as sentiment toward a current health topic (e.g., vaccines).[11] Myslín et al. analyzed tweets to examine sentiment toward various tobacco products and found that hookah and electronic cigarettes were characterized by more positive sentiment than references to traditional tobacco products and general references to smoking behavior.[12] Liking or following alcohol marketing social media pages has been found to be associated with early age at first alcohol consumption and heavier alcohol consumption among youths.[13]

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

ABOUT THE AUTHORS

Quynh C. Nguyen is with the Department of Epidemiology and Biostatistics, University of Maryland School of Public Health, College Park. Hsien-wen Meng and Geoffrey Loomis are with the Department of Health, Kinesiology, and Recreation; University of Utah College of Health; Salt Lake City. Matt McCullough and Dapeng Li are with the Department of Geography, University of Utah. Debjyoti Paul, Suraj Kath, and Feifei Li are with the School of Computing, University of Utah. Elaine O. Nsoesie is with Institute for Health Metrics and Evaluation, University of Washington, Seattle. Ming Wen is with the Department of Sociology, University of Utah. Ken R. Smith is with the Department of Family Consumer Studies, University of Utah.

Correspondence should be sent to Quynh C. Nguyen, Department of Epidemiology and Biostatistics, University of Maryland School of Public Health, 255 Campus Drive, College Park, MD 20742 (e-mail: qtnguyen@umd.edu). Reprints can be ordered at http://www.ajph.org by clicking the "Reprints" link.

This article was accepted July 2, 2017.

doi: 10.2105/AJPH.2017.303993

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 enabled 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.[14,15] More details of our methodology can be found in Nguyen et al.[16]

Sentiment Analysis

We used a Java-based package for natural language processing, MAchine Learning for LanguagE Toolkit (MALLET; McCallum, Amherst, MA), for sentiment analysis. We used labeled tweets to train the maximum entropy text classifier in MALLET.[17] 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),[18] Sanders Analytics (n = 5513 tweets),[19] and Kaggle (n = 7086 tweets).[20]

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."[15,16] 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.[21,22] 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

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 1—Descriptive Statistics, County Level: Contiguous United States, April 2015–March 2016

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

[a]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.

[b]Data sources for health outcomes: 2011–2013 Centers for Disease Control and Prevention WONDER (Wide-Ranging Online Data for Epidemiologic Research) mortality data; 2011–2014 Behavioral Risk Factor Surveillance System on adults aged 20 years and older.

[c]We defined premature mortality as deaths occurring before age 75 years.

[d]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.

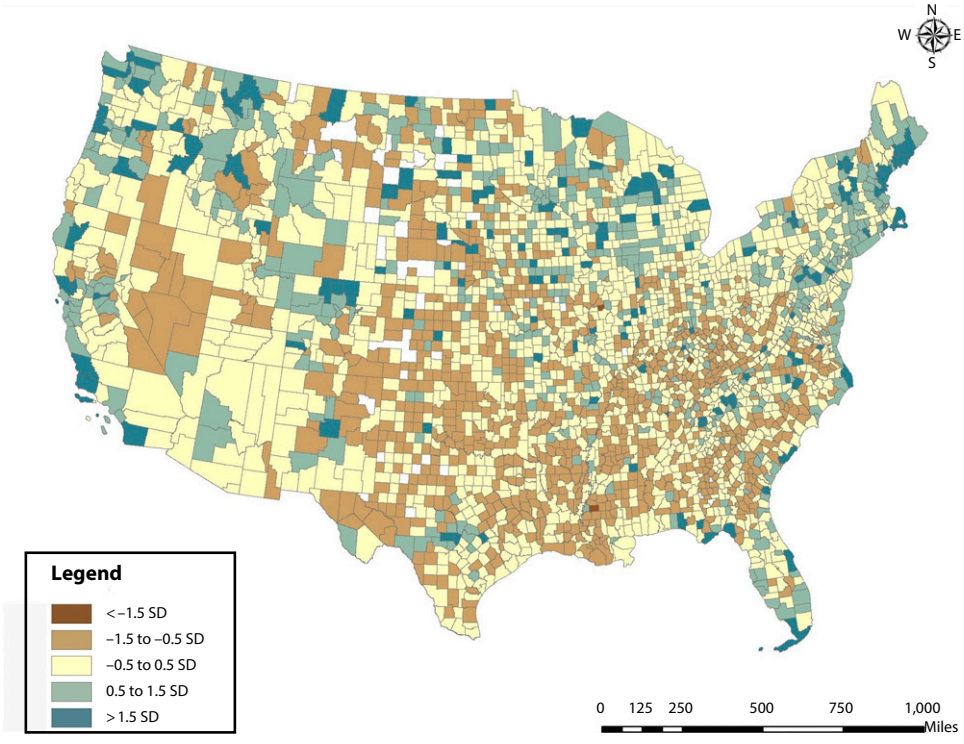
calories per 100 grams. Tweets about healthy foods were happier than those about fast food (25% vs 17%). On average, 2% of tweets were about physical activity and less than 1% of tweets mentioned alcohol use. Tweets about alcohol use were slightly happier than those about physical activity or healthy 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). The proportion of happy tweets was highest in Montana, Tennessee, Utah, New Hampshire, Arkansas, Maine, Colorado, and New York (Figure H, 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



Note. Prevalence of Twitter food mentions was estimated by using a dictionary of 1430 popular foods. Values were standardized to have a mean of 0 and standard deviation of 1. Negative values indicate below national average values. Positive values indicate above national average values. County map of Twitter characteristics were unadjusted for county characteristics.

FIGURE 1—National Distribution of Twitter Food Mentions, County Level: Contiguous United States, April 2015–March 2016

TABLE 2—Twitter Characteristics as Predictors of Health Outcomes, County Level: Contiguous United States, 2011–2013

County-Level Twitter Predictors[a]	No.	Percentage Obesity, B (95% CI)[b]	Percentage Physical Inactivity, B (95% CI)[b]
Food tweets			
Third tertile (highest)	3057	−2.49 (−3.23, −1.76)	−3.62 (−4.44, −2.80)
Second tertile		−0.61 (−1.07, −0.15)	−1.46 (−2.06, −0.86)
Physical activity tweets			
Third tertile (highest)	3054	−2.40 (−3.33, −1.47)	−2.97 (−3.85, −2.08)
Second tertile		−1.01 (−1.50, −0.52)	−1.39 (−1.85, −0.93)
Happy tweets			
Third tertile (highest)	3117	−2.23 (−3.15, −1.31)	−1.97 (−3.09, −0.86)
Second tertile		−0.79 (−1.31, −0.28)	−0.68 (−1.28, −0.07)

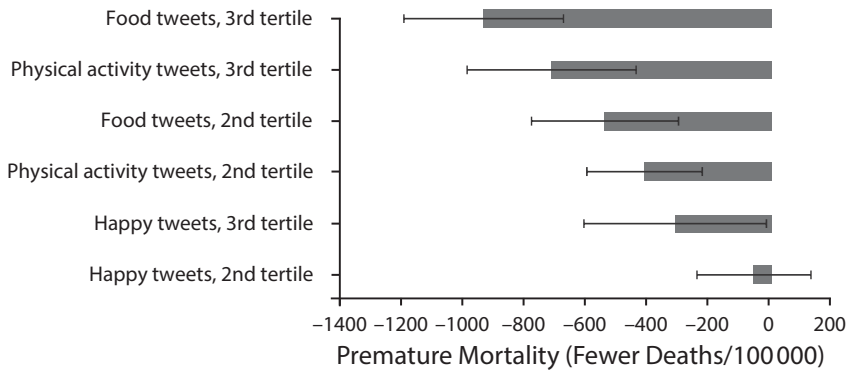
Note. CI = confidence interval. Data sources for health outcomes: 2011–2013 National Center for Health Statistics for prevalence of obesity and leisure-time physical activity age-adjusted to 2000 US population.

[a]County-level predictors of happiness, food, and physical activity were derived from 80 million tweets, 4 million food tweets, and 1.5 million physical activity tweets, respectively.

[b]Adjusted linear regression models were run for each outcome separately. Models controlled for county-level demographics: median age, % non-Hispanic White, median household income. Standard errors accounted for clustering of county values at the state level. Twitter characteristics were categorized into tertiles, with the lowest tertile serving as the referent group.

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).



Note. Twitter characteristics were categorized into tertiles, with the lowest tertile serving as the referent group. Data source for health outcomes: 2011–2013 Centers for Disease Control and Prevention WONDER (Wide-Ranging Online Data for Epidemiologic Research) mortality data. Premature deaths are deaths occurring before age 75 years. Adjusted linear regression models controlled for county-level demographics: median age, percentage non-Hispanic White, and median household income. Standard errors accounted for clustering of county values within a state. Coefficients and 95% confidence intervals are displayed in the figure.

FIGURE 2—Twitter Characteristics as Predictors of Premature Mortality, County Level: Contiguous United States, April 2015–March 2016

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.[24] 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.[25]

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.[26] 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, 9967 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 nonrepresentativeness 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. AJPH

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. Kath 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. Nsoesie, M. Wen, 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

1. Bostic RW, Thornton RL, Rudd EC, Sternthal MJ. Health in all policies: the role of the US Department of Housing and Urban Development and present and future challenges. *Health Aff (Millwood)*. 2012;31(9):2130–2137.

2. Marmot M, Friel S, Bell R, Houweling TA, Taylor S; Commission on Social Determinants of Health. Closing the gap in a generation: health equity through action on

the social determinants of health. *Lancet*. 2008;372(9650): 1661–1669.

3. Blas E, Gilson L, Kelly MP, et al. Addressing social determinants of health inequities: what can the state and civil society do? *Lancet*. 2008;372(9650): 1684–1689.

4. Diez Roux AV, Merkin SS, Arnett D, et al. Neighborhood of residence and incidence of coronary heart disease. *N Engl J Med*. 2001;345(2):99–106.

5. Jia H, Moriarty DG, Kanarek N. County-level social environment determinants of health-related quality of life among US adults: a multilevel analysis. *J Community Health*. 2009;34(5):430–439.

6. Black NC. An ecological approach to understanding adult obesity prevalence in the United States: a county-level analysis using geographically weighted regression. *Appl Spat Anal Policy*. 2014;7(3):283–299.

7. Mehta NK, Chang VW. Weight status and restaurant availability a multilevel analysis. *Am J Prev Med*. 2008; 34(2):127–133.

8. Nagel AC, Tsou M-H, Spitzberg BH, et al. The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets. *J Med Internet Res*. 2013;15(10):e237.

9. Harris JK, Hawkins JB, Nguyen L, et al. Using Twitter to identify and respond to food poisoning in real time: the Food Safety STL project. *J Public Health Manag Pract*. 2017; Epub ahead of print.

10. Imran M, Castillo C, Lucas J, Meier P, Vieweg S. AIDR: Artificial intelligence for disaster response. In: *Proceedings of the 23rd International Conference on World Wide Web*. New York, NY: Association for Computing Machinery; 2014: 159–162.

11. Bahk CY, Cumming M, Paushter L, Madoff LC, Thomson A, Brownstein JS. Publicly available online tool facilitates real-time monitoring of vaccine conversations and sentiments. *Health Aff (Millwood)*. 2016;35(2):341–347.

12. Myslín M, Zhu S-H, Chapman W, Conway M. Using Twitter to examine smoking behavior and perceptions of emerging tobacco products. *J Med Internet Res*. 2013; 15(8):e174.

13. Carrotte ER, Dietze PM, Wright CJ, Lim MS. Who "likes" alcohol? Young Australians' engagement with alcohol marketing via social media and related alcohol consumption patterns. *Aust N Z J Public Health*. 2016; 40(5):474–479.

14. Guttman A. R-trees: a dynamic index structure for spatial searching. In: *Proceedings of the 1984 ACM SIGMOD International Conference on Management of Data*. New York, NY: Association for Computing Machinery; 1984;14(2):47–57.

15. Nguyen QC, Kath S, Meng H-W, et al. Leveraging geotagged Twitter data to examine neighborhood happiness, diet, and physical activity. *Appl Geogr*. 2016;73: 77–88.

16. Nguyen QC, Li D, Meng H-W, et al. Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. *JMIR Public Health Surveill*. 2016;2(2):e158.

17. Nigam K, Lafferty J, McCallum A. Using maximum entropy for text classification. In: *IJCAI-99 Workshop on Machine Learning for Information Filtering*. Stockholm, Sweden: International Joint Conference on Artificial Intelligence; 1999: 61–67.

18. Sentiment140. For academics. 2009. Available at: http://help.sentiment140.com/for-students. Accessed March 2, 2016.

19. Sanders Analytics. Twitter sentiment corpus. 2011. Available at: http://www.sananalytics.com/lab/twitter-sentiment. Accessed March 3, 2016.

20. Kaggle in Class. Sentiment classification. 2011. Available at: https://inclass.kaggle.com/c/si650winter11. Accessed March 3, 2016.

21. Ainsworth BE, Haskell WL, Herrmann SD, et al. 2011 compendium of physical activities: a second update of codes and MET values. *Med Sci Sports Exerc*. 2011;43(8): 1575–1581.

22. Zhang N, Campo S, Janz FK, et al. Electronic word of mouth on Twitter about physical activity in the United States: exploratory infodemiology study. *J Med Internet Res*. 2013;15(11):e261.

23. Nguyen QC, McCullough M, Pham M. HashtagHealth: a social media big data resource for neighborhood effects funded by NIH's Big Data to Knowledge (BD2K) initiative. 2017. Available at: https://hashtaghealth.github.io/geoportal/start.html. Accessed August 16, 2017.

24. Abbar S, Mejova Y, Weber I. You tweet what you eat: studying food consumption through Twitter. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery; 2015: 3197–3206.

25. Artiga S, Damico A. Health and health coverage in the South: a data update. Menlo Park, CA: Henry J. Kaiser Family Foundation; 2016.

26. França LR, Dautzenberg B, Reynaud M. Heavy episodic drinking and alcohol consumption in French colleges: the role of perceived social norms. *Alcohol Clin Exp Res*. 2010;34(1):164–174.

27. Rosenquist JN, Murabito J, Fowler JH, Christakis NA. The spread of alcohol consumption behavior in a large social network. *Ann Intern Med*. 2010;152(7): 426–433.

28. US Department of Transportation, National Highway Traffic Safety Administration. Traffic safety facts 2014 data: alcohol-impaired driving. Washington, DC: National Highway Traffic Safety Administration; 2015.

29. Koopmans TA, Geleijnse JM, Zitman FG, Giltay EJ. Effects of happiness on all-cause mortality during 15 years of follow-up: The Arnhem Elderly Study. *J Happiness Stud*. 2010;11(1):113–124.

30. Anton SD, Miller PM. Do negative emotions predict alcohol consumption, saturated fat intake, and physical activity in older adults? *Behav Modif*. 2005;29(4):677–688.

31. Physical Activity Guidelines Advisory Committee. *Physical Activity Guidelines Advisory Committee Report, 2008*. Washington, DC: US Department of Health and Human Services; 2008.

32. Morstatter F, Pfeffer J, Liu H, Carley KM. Is the sample good enough? Comparing data from Twitter's streaming API with Twitter's Firehose. In: *Proceedings for the 7th International AAAI Conference on Web Blogs and Social Media*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence; 2013. Available at: https://arxiv.org/abs/1306.5204. Accessed August 16, 2017.

33. Sloan L, Morgan J. Who Tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. *PLoS One*. 2015;10(11):e0142209.

34. Duggan M. *The Demographics of Social Media Users*. Washington, DC: Pew Research Center; 2015.

35. Wierzbicka A. "Happiness" in cross-linguistic and cross-cultural perspective. *Daedalus*. 2004;133(2):34–43.

36. Matsumoto D, Yoo SH, Fontaine J. Mapping expressive differences around the world the relationship between emotional display rules and individualism versus collectivism. *J Cross Cult Psychol*. 2008;39(1):55–74.