



ARTICLE

<https://doi.org/10.1057/s41599-019-0314-x>

OPEN

Use of social media, search queries, and demographic data to assess obesity prevalence in the United States

Nina Cesare^{1,2}, Pallavi Dwivedi³, Quynh C. Nguyen³ & Elaine O. Nsoesie^{1,2}

ABSTRACT Obesity is a global epidemic affecting millions. Implementation of interventions to curb obesity rates requires timely surveillance. In this study, we estimated sex-specific obesity prevalence using social media, search queries, demographics and built environment variables. We collected 3,817,125 and 1,382,284 geolocated tweets on food and exercise respectively, from Twitter's streaming API from April 2015 to March 2016. We also obtained searches related to physical activity and diet from Google Search Trends for the same time period. Next, we inferred the gender of Twitter users using machine learning methods and applied mixed-effects state-level linear regression models to estimate obesity prevalence. We observed differences in discussions of physical activity and foods, with males reporting higher intensity physical activities and lower caloric foods across 40 and 48 states, respectively. In addition, counties with the highest percentage of exercise and food tweets had lower male and female obesity prevalence. Lastly, our models separately captured overall male and female spatial trends in obesity prevalence. The average correlation between actual and estimated obesity prevalence was 0.797(95% CI, 0.796, 0.798) and 0.830 (95% CI, 0.830, 0.831) for males and females, respectively. Social media can provide timely community-level data on health information seeking and changes in behaviors, sentiments and norms. Social media data can also be combined with other data types such as, demographics, built environment variables, diet and physical activity indicators from other digital sources (e.g., mobile applications and wearables) to monitor health behaviors at different geographic scales, and to supplement delayed estimates from traditional surveillance systems.

¹Department of Global Health, Boston University School of Public Health, Boston, Massachusetts, USA. ²Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington, USA. ³Department of Epidemiology and Biostatistics, University of Maryland School of Public Health, College Park, Maryland, USA. Correspondence and requests for materials should be addressed to E.O.N. (email: onelaine@bu.edu)

Background

The rate of obesity in both children and adults in the United States has increased significantly since the 1980s (Dwyer-Lindgren et al., 2013; Fryar et al., 2016; Segal et al., 2017). In 2017, the State of Obesity project estimated that adult obesity prevalence across U.S. states ranged from 22.3 to 37.7 percent (Segal et al., 2017). This increase in obesity prevalence is due to a complex interplay of biological, structural and individual factors (Hill and Peters, 1998; Nelson et al., 2006; Papas et al., 2007; Ogden et al., 2010). Factors such as public safety, socioeconomic status, and the neighborhood built environment may impact access to recreational facilities, and fresh, healthy foods (Freedman et al., 2002; Giles-Corti et al., 2003; Hill et al., 2003; Ellaway et al., 2005; Gordon-Larsen et al., 2006; Lopez-Zetina et al., 2006; Mobley et al., 2006; Bennett et al., 2007; Papas et al., 2007; Casagrande et al., 2009; Maharana and Nsoesie, 2018). An individual's social environment can also influence health behaviors (such as, poor diet and physical inactivity) that are considered risk factors for obesity (Christakis and Fowler, 2007; McFerran et al., 2009; Yakusheva et al., 2011).

Timely surveillance of changes in physical activity prevalence and dietary habits would be valuable for understanding the association between these factors and obesity prevalence in communities. One approach to surveillance is the use of data from digital sources, which are useful for estimating various health quantities by monitoring the seeking and sharing of health information, as well as changes in behaviors, sentiments or norms (Cesare et al., 2019; Olson et al., 2013; Yuan et al., 2013; Nsoesie et al., 2014; Nsoesie et al., 2015; Nguyen et al., 2016; Nguyen et al., 2017). Specifically, data from Twitter—a social media platform—have been shown to provide insights into trends and disparities in health behaviors including, diet and physical activity (Griffin and Jiao, 2015; Jestico et al., 2016; Nguyen et al., 2017; Cesare et al., 2019; Torres et al., 2018). One study showed a correlation between Facebook likes of sedentary activities and obesity prevalence (Chunara et al., 2013). Another study found a negative association between the volume of food and exercise-related tweets and obesity in U.S. counties (Nguyen et al., 2017). Neither of these studies included sex-specific analysis. Sex-specific analysis is important because obesity prevalence varies by sex in the U.S. with higher prevalence reported for females than males in recent years (Ogden et al., 2014; Flegal et al., 2016; Tauqeer et al., 2018).

In this study, we aimed to assess the association between obesity prevalence estimated by the Centers for Disease Control and Prevention (CDC) and various food and exercise variables from social media (i.e., Twitter) and search queries (i.e., Google Search Trends) for males and females separately. We also demonstrated that integrating data from the aforementioned Internet sources with demographics and built environment variables could be useful for estimating obesity prevalence in U.S. counties by sex.

Methods

Sex-specific, county-level obesity estimates. Age-adjusted obesity estimates for U.S. counties were downloaded from the CDC. These estimates were derived by applying a small area estimation technique to data from the Behavioral Risk Factors and Surveillance System (BRFSS)—a telephone survey on health behaviors related to chronic diseases, injury, and preventable infectious diseases for the non-institutionalized adult U.S. population (Malec et al., 1997; Centers for Disease Control and Prevention, 2018a).

The most recent county-level obesity estimates by sex from the CDC were based on the 2013 BRFSS survey. To align the CDC data with the Twitter data which was collected between April

2015 and March 2016, we used linear autoregressive models to forecast 2015 obesity prevalence. Our model used estimates from previous years to estimate 2015 obesity prevalence. The model R^2 (i.e., coefficient of determination) was 82.73% and 82.73% for males and females, respectively. While the State of Obesity project reported an increase in obesity prevalence for all but seven states between 2013 and 2016, this increase was only significant for three states: Alabama, Michigan and Nebraska (see SI Fig. 1) (Segal et al., 2017). We used both the 2013 obesity estimates and the 2015 projections in our analysis.

Social media data. We used Twitter's public streaming application programming interface (API) to collect a random 1% of publicly available geotagged (i.e., including a latitude and longitude) tweets (Twitter Developers, 2014). We collected tweets related to food and exercise based on a set of 1430 keywords from the U.S. Department of Agriculture's National Nutrient Database (United States Department of Agriculture, 2014) and 376 keywords from common fitness questionnaires and apps (Ainsworth et al., 2000; Zhang et al., 2013; Nguyen et al., 2016), respectively. The keywords consisted of popular foods, beverages and fast food restaurants. Fruits, vegetables, nuts, and lean proteins (e.g., fish, chicken, and turkey) were labeled as 'healthy foods'. Since Twitter allows access to information on a variety of physical activities—not just those associated with transportation or planned workouts—the keywords for physical activities consisted of sports, recreational activities, household chores, and gym-related activities. The resulting data consisted of 79,848,992 million postings from 603,363 users and was collected from April 2015 to March 2016. Each tweet was mapped to a county based on the latitude and longitude using geospatial shapefiles. The data for this project were collected as part of a larger initiative to assess small area health trends using digital data. A description of this larger project can be found at <http://hashtaghealth.github.io>. Data collection and processing are described in the proceeding paragraphs, and in further detail in the cited papers (Nguyen et al., 2016; Nguyen et al., 2017).

Social media data processing. The data were cleaned to exclude duplicates, outliers (i.e., users whose tweets represented greater than 1% of tweets), job postings, and tweets falling outside of the contiguous United States. The Maximum Entropy text classifier in the Machine Learning for Language Toolkit (MALLET) (McCallum, 2002) was used to classify tweet sentiment between zero and one, with one indicating the strongest positive sentiment. This classification was carried out with the broader project aims of assessing happiness in U.S. counties and evaluating its association with various health outcomes including, premature mortality, diabetes and obesity. The classifier was rigorously trained using existing and publicly available datasets from Sentiment140 (Sentiment140, 2009), Sanders Analytics (Sanders Analytics, 2011), and Kaggle (Kaggle. Sentiment classification, 2011). While MALLET is not the only sentiment toolkit available, we found that it outperformed a bag-of-words approach, Sentiment140, and standard supervised machine learning classifiers. When compared to 500 manually labeled tweets, the accuracy of our sentiment scores was 77%.

Of the ~80 million general topic tweets collected, a total of 3,817,125 tweets were identified as containing at least one food related keyword. There was a median of 12 food tweets per user. We used a text matching algorithm to identify food versus non-food tweets. This algorithm iteratively identified two-word foods (e.g., orange chicken) and then went through the data again to identify one-word foods (e.g., taco). To assess the performance we applied the algorithm to 2500 manually labeled tweets (2000

food-related and 500 non-food related). The accuracy and F1-score (the harmonic average of the precision and recall; 1 is the best possible score) were 0.83 and 0.86, respectively. Precision is defined as the ratio of true positive classifications to all positive cases, and recall is defined as the ratio of true positive classifications to all correctly predicted cases. We compared our approach to several supervised learning approaches (i.e., feed forward neural network (FFNN), support vector machines (SVM), gradient boosting and fastText (Joulin et al., 2016)) and found that our approach performed better.

The caloric density defined as calories per 100 g was estimated for each food based on data from the USDA. The caloric density for each tweet was computed by summing the associated calories for each food mentioned in the tweet. The prevalent sentiment of each food tweet was also ascertained using the previously described sentiment analysis process.

A total of 1,382,284 tweets contained at least one physical activity keyword. There was a median of five tweets per user. To identify exercise tweets, we used a keyword matching algorithm that removed popular phrases that do not denote physical activity (e.g., ‘walk away’ or ‘running late’), phrases associated with pop culture (e.g., ‘Walking Dead’), and terms that denote watching rather than participating in exercise (e.g., “attend” and “watch”). For team sports, we only retained tweets that contained the words play/playing/played in conjunction with the activity. To assess the performance of this text-matching algorithm, 2500 tweets were manually labeled (2000 exercise-related and 500 non-exercise-related). The accuracy was 85% and the F1-score was 0.90. Exercise intensity (hereafter referred to as, calories burned) was quantified using the metabolic equivalent associated with the performance of each activity for a 30 min duration by a 155-pound individual, the average weight of an American adult (Ainsworth et al., 2000; Harvard Health Publications, 2015). For additional details on data processing, see (Nguyen et al., 2017).

Demographic inference of social media users. Data from Twitter does not include demographic characteristics of users. To address this limitation, we developed a scalable and efficient ensemble approach for inferring gender by combining predictions from three previously proposed methods that focus only on the metadata available on users’ profile (Burger et al., 2011; Mislove et al., 2011; Longley et al., 2015; Mueller and Stumme, 2016). The three approaches included, method (1), Twitter users’ first names were matched to data from the U.S. Social Security Administration (Longley et al., 2015) (this captured approximately 60% of Twitter names); method (2), we used word and character n-grams from users’ names and a Support Vector Machine (SVM) classifier with a linear kernel (Burger et al., 2011); and method (3), we applied a decision tree classifier to the linguistic structure of users’ names—including the count of syllables, vowels, consonants, boubas (round) and kiki (sharp) vowels and consonants (Maurer et al., 2006; Nielsen and Rendall, 2011), and whether or not the last character was a vowel. We combined the prediction from all three classification methods using an ensemble approach—weighted stacked logistic regression (Wolpert, 1992). The ensemble classifier achieved an accuracy of 0.827, recall of 0.852, and an F1-score of 0.837. It outperformed methods (2) and (3), and unlike method (1), it captured all users with alphanumeric names. For a complete description of the gender inference process, refer to the following references (Cesare et al., 2017a; Cesare et al., 2017b). We used bootstrapping to quantify uncertainty in our inference as described in the Error section.

We applied the ensemble classifier to infer the gender for each user in the previously described food and physical activity Twitter datasets. We then generated county-level sex-specific variables for

food and physical activity including, the proportion of food, healthy food and fast food tweets, sentiment towards food, sentiment towards physical activity, proportion of physical activity tweets, calories consumed and calories burned.

Google search trends (GST). We used Google Trends (<https://trends.google.com/trends/>) to obtain state-level searches for the phrases: *fitness center*, *fast food*, *weight loss*, *organic food* and *grocery store*. We used state-level data because county-level data was unavailable. After examining correlations between these variables, we selected the terms—*fitness center*, *fast food*, and *grocery store*—to avoid multicollinearity. The data was scaled by Google to have a maximum of one hundred, such that states with the highest volume of searches had a value of one hundred.

Statistical analysis. To assess the association between postings on Twitter, and survey estimates of obesity prevalence at the county level, we fitted separate linear mixed effects regression models with a varying-intercept group effect at the state level to account for variations among states for males and females. The model can be specified as follows:

$$Y_{ij} = X_{ij}\beta + U_{ij}\gamma_i + \epsilon_i, \quad i = 1, \dots, m \text{ and } j = 1, \dots, n_i$$

$$\gamma_i \sim N_q(0, D) \quad \epsilon_i \sim N_{n_i}(0, \Sigma_i)$$

Y_{ij} = response of j th county of state i

m = number of states

n_i = number of counties in state i

X_{ij} = covariate vector of j th county of state i for fixed effects

β = fixed effects parameter

U_{ij} = covariate vector of j th county of state i for random effects

γ_i = random effect parameter

To estimate obesity prevalence, we compared four models: (1) a model that included only digital data variables, (2) a model that included demographic and built environment variables, (3) a model that included Twitter and demographic variables and (4) a model that included Twitter, demographic and built environment variables. These models had a better goodness of fit than simple linear regression models. We used the lmer package in R (R Core Team, 2013).

The sociodemographic measures included in our analysis were household income, healthy food availability, and neighborhood safety; measures previously associated with obesity prevalence (Nelson et al., 2006; Mattes and Foster, 2014; Nesbit et al., 2014; Cooksey-Stowers et al., 2017). Specifically, we obtained measures of median household income, median county age by sex, percent non-Hispanic black and percent Hispanic by sex from the 2015 5-year American Community Survey (United States Census Bureau, 2015). We also included county-level measures of the built environment that may impact access to food and exercise. These measures included, the count of fast food restaurants and grocery stores per 1000 residents from the U.S. Department of Agriculture (USDA) (United States Department of Agriculture, 2018) and violent crime rate, the percent of residents who have access to exercise space, the percent of residents who are food insecure, the percent of residents who drive alone to work, and the violent crime rate from the County Health Rankings and Roadmaps project (County Health Rankings and Roadmaps, 2016).

All percentages and sentiment scores were scaled to have values between zero and one. The proportion of exercise, fast food, food and healthy food tweets were divided into three equally spaced groupings (hereafter referred to as, tertile). Income was quantified in thousands of dollars.

We used five-fold cross validation and assessed model goodness of fit using the coefficient of determination (R^2). We also used logistic regression to investigate male and female differences in caloric intake and calories burned.

Error. A major limitation of many machine learning methods is a lack of uncertainty quantification. We quantified the uncertainty in our overall methodology using bootstrapping (Efron and Tibshirani, 1986). We created 100 datafiles in which we (a) randomly sampled 16% of the users, since our accuracy rate in classifying gender was 84% and (b) randomly assigned these users a male/female label. We used 5-fold cross validation to generate out-of-sample predictions for each dataset. We reported the range of estimated obesity prevalence and the correlation between our model estimates with data from the CDC, by sex.

Ethics approval and consent to participate. This study was deemed exempt by the Institutional Review Board at the University of Washington, where the study was started.

Consent for publication. All authors have consented to the publication of this paper.

Results

Summary of data. The Twitter data consisted of 1,382,284 physical activity tweets (481,146 users) and 3,817,125 food tweets (775,002 users). See Table 1 for a sample of food and physical activity tweets by sex. About 15.8% (604,907) of the food tweets were classified as healthy and 9.17% (350,024) were classified as fast foods. Note that not all tweets received a healthy or fast food classification. These food postings originated from 3004 and 2985 unique counties separately for males and females. Similarly, the

physical activity postings were from 2992 and 2932 counties, respectively for males and females. The U.S. has 3007 counties.

Table 1: Tweets are intended to provide context regarding how individuals tweet about food and physical activity. Note that the text content of tweets has been slightly modified to preserve anonymity. Specific handles and URLs have been removed. For instance, “Out on a run” might be changed to “Off on a run”.

Counties with the highest percentage of exercise and healthy food tweets also had a higher median age, suggesting older adults are more likely to post about exercise and healthy foods (SI Fig. 2). In contrast, counties with a lower median age had a higher percentage of fast food tweets. Counties with a higher number of fast food restaurants per 1,000 persons, also had the highest percentage of fast food tweets suggesting the online social environment may reflect the built environment (Fig. 1a). Furthermore, counties with the highest percentage of food insecure households also had the lowest percentage of exercise and food tweets (Fig. 1b). Conversely, counties with the highest number of grocery stores per 1,000 residents had the highest proportion of exercise and food tweets (SI Fig. 3). Lowest overall exercise, food, healthy and fast food tweets were noted for counties with lowest socioeconomic status (SES). These observations were consistent for males and females.

In contrast, males were more likely to report higher intensity physical activities (SI Fig. 4) and foods associated with lower caloric intake (SI Fig. 5) across 40 and 48 states, respectively. Specifically, males were 0.924 ($p < 0.001$) and 0.8254 ($p < 0.001$) times less likely to mention foods that could be clearly classified as fast and healthy foods, respectively. The contrast between males and females was more pronounced for postings of foods than physical activity.

Associations between data from GST, Twitter, and obesity prevalence from the CDC. There were differences in obesity

Table 1 A random sample of exercise and food tweets, by sex

Exercise tweets: female

“I just finished running 4.50 miles in 1h:00 m:02 s with #endomondo #endorphins [url]”
 “1st Saturday workout of 2016. Weird to say haha... #teamcoconutz take over. no filters on this [url]”
 “@username celebrities, Alzheimer’s disease is nothing to ignore. Help me reach my goal for my walk [url]”
 “Out on a run... (@ Ferry building in San Francisco, CA) [url]”
 “It’s golf time @ walnut creek golf club and banquet center [url]”
 “I just finished walking 3.24 miles in 55 m:38 s with #endomondo #endorphins [url]”

Exercise tweets: male

“Good times at fsbm mission tournament. @ Lake Louise golf club [url]”
 “I’m at Madison swim academy in [city location] [url]”
 “I miss this this right about now #skiing #mountains #beavercreek #snow #snowboarding [url]”
 [url] Everything’s ok -- on the TMJ thru hike!”
 “Such a beautiful evening to play golf with @username @ Wilson and Harding golf course [url]”
 “You have to love dancing to stick with it. It gives nothing back, there are no manuscripts to store [url]”

Food tweets: female

Tonight’s wisdom from Panda Express... @ H&R House [url]
 Mom date with hotchocolate + whipped cream @ Starbucks [url]
 Today’s “short stop to say hey” turned into a 3 hour bbq at my grandpas
 #DrinkGoodDoGood @nakedjuice I love apples [url]
 When your curry wrap photobombs your fries [emoji]. To summarize, my chicken tikka at mymasalawrap [url] Shipping out to a new, happyhome!
 #Unicorn #LePin #wine @ The Wine Vault [url]
 When your curry wrap photobombs your fries [emoji]. To summarize, my chicken tikka at mymasalawrap [url]

Food tweets: male

Last night’s #pizza @ Perrot’s Chicago Pizza [url]
 Banana Nutella Pop Tart with Chocolate Pop Rocks. #Perfection #SuperbaFoodAndBread #Superba #Venice [url] I’m at Jim’s Pizza & Wings - @username in Orlando, FL [url]
 My quest for real bacon has brought me here <http://t.co/bslqxupF7M>
 Classic #margarita and #lamb #tacos. A wonderful end to a great weekend.& [url]
 So Chipotle it is where the total remaining population decided to eat.
 I’m at Jim’s Pizza & Wings - @username in Orlando, FL [url] Really enjoyable beer. - Drinking a Hop Lab Simcoe Wheat IPA by @kanebrewing @ Sparta [emoji] [url]

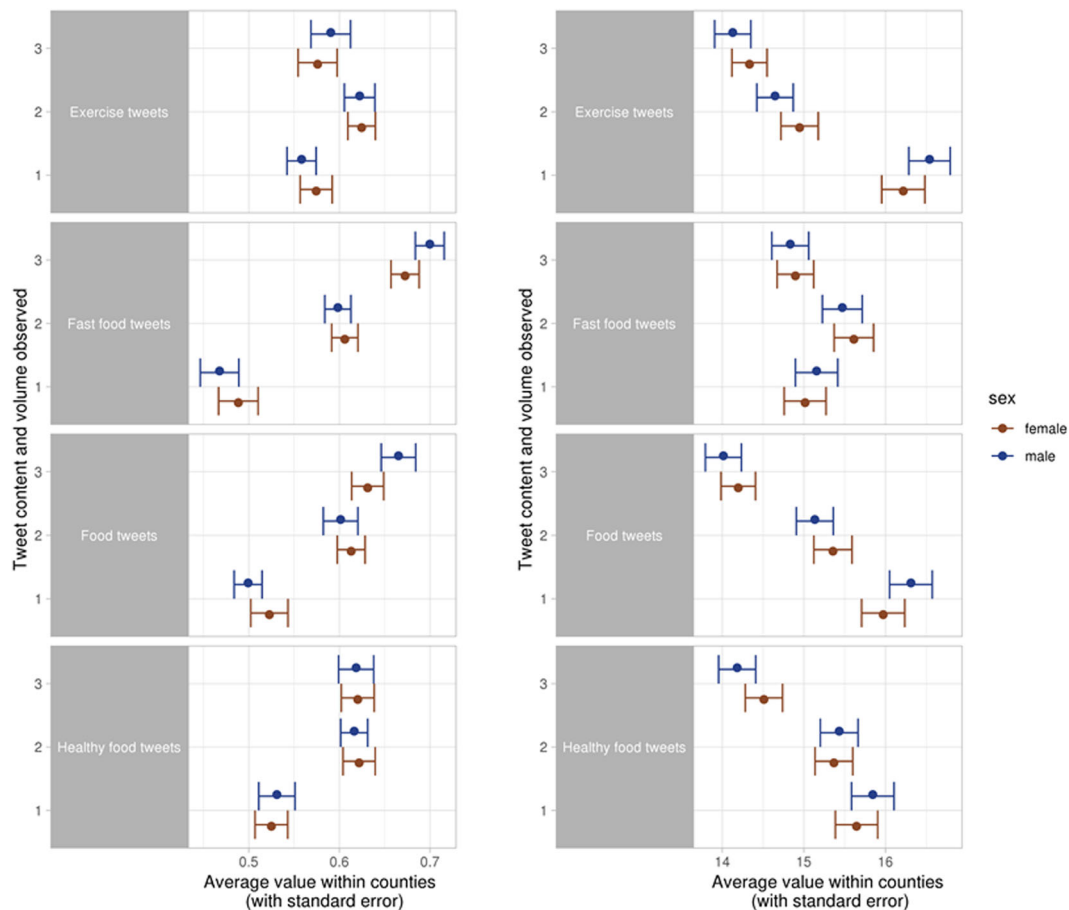


Fig. 1 Relationship between density of fast food restaurants (left) and percent food insecure (right), and proportion of healthy food, food, fast food and exercise tweets across counties. The proportion of exercise, fast food, food and healthy food tweets are divided into three tertile (i.e., equally spaced groupings): 1 (low) to 3 (high). Overall, the observations for males and females are consistent with some minor differences in the mean and range

prevalence across counties with lower overall prevalence noted for males compared to females. The observations that follow were noted for both males and females (Tables 2 and 3). Counties with more positive sentiment towards food and exercise had lower obesity prevalence. Similarly, counties with the highest percent of exercise, food, healthy food and fast food tweets had lower obesity prevalence. These relationships remained after controlling for demographics and built environment variables. However, counties in the middle tertile for fast food tweets had higher obesity prevalence, and this relationship remained after controlling for demographic and built environment variables. The coefficients for the Google search trends were much smaller when compared to the Twitter data variables, and not statistically significant (see SI Table 1). The main differences between the models for males and females were in the size of the model coefficients, and the models for females had a slightly better goodness of fit. The R^2 ranged from 0.56 to 0.67, and 0.57 to 0.61 for females and males, separately.

Estimation of obesity prevalence using GST, Twitter, and demographics data. Models that combined GST, Twitter, socio-demographic and built environment variables explained 69.0% and 61.5% of the variation in obesity prevalence for females and males, separately. The average R^2 from our error analysis was 0.624 (95% CI 0.624, 0.625) and 0.553 (95% CI 0.552, 0.553), for females and males, respectively. These values are lower than those obtained in Tables 2 and 3, suggesting that our gender

classification approach is more reliable than a random assignment of gender to a subset of the population.

Differences between our model estimates and county-level obesity prevalence from the CDC were consistent across the US with some exceptions (see Fig. 2). We also observed similar trends when we used the projected obesity prevalence for 2015 (SI Table 2). See SI Fig. 6 for the difference between 2013 CDC-estimated obesity and our model-estimated obesity prevalence for males and females. Our models captured the overall linear trend in obesity prevalence for both males and females. The average correlation between actual and predicted values for females and males were separately 0.830 (95% CI 0.830, 0.831) and 0.797 (95% CI 0.796, 0.798) based on out-of-sample estimates. The R^2 values suggest these data may be better suited for estimating obesity prevalence for females. Furthermore, our obesity prevalence estimates are aided by the inclusion of sociodemographic and built environment variables (Tables 2 and 3).

Discussion

Digital technologies can be used for monitoring behavioral factors related to obesity, such as decisions on diet and exercise (Said and Bellogín, 2014; Griffin and Jiao, 2015; De Choudhury et al., 2016; Jestic et al., 2016; Torres et al., 2018). Here, we show that sex-specific data can provide valuable insights into how males and females discuss exercise and food on social media. We also demonstrate that these data could be useful for monitoring sex-

Table 3 Coefficients for linear mixed-effects models for male obesity prevalence

	Twitter variables	Built environment and demographic variables	Twitter and demographic variables	Twitter, demographic and built environment variables
Tweet sentiment toward food	-0.011***(0.004)		-0.009**(0.004)	-0.007*(0.004)
Percent of tweets food related (2nd tertile)	0.007(0.148)		0.088(0.141)	0.144(0.138)
Percent of tweets food related (3rd tertile)	-1.192***(0.181)		-0.909***(0.173)	-0.651***(0.172)
Percent of tweets fast food related (2nd tertile)	0.156(0.135)		0.326**(0.131)	0.435***(0.134)
Percent of tweets fast food related (3rd tertile)	-0.791***(0.148)		-0.378**(0.15)	-0.04(0.156)
Percent of tweets healthy food related (2nd tertile)	0.207(0.142)		0.151(0.135)	0.179(0.133)
Percent of tweets healthy food related (3rd tertile)	-0.325**(0.162)		-0.232(0.154)	-0.109(0.152)
Tweet sentiment toward exercise	-0.007**(0.003)		-0.006**(0.003)	-0.002(0.003)
Percent of tweets exercise related (2nd tertile)	-0.480***(0.138)		-0.296**(0.132)	-0.186(0.129)
Percent of tweets exercise related (3rd tertile)	-1.100***(0.145)		-0.787***(0.141)	-0.541***(0.139)
Median age within male population		-0.098***(0.012)	-0.044***(0.012)	-0.079***(0.013)
Percent non-Hispanic black within male population		0.027***(0.008)	0.007(0.005)	0.026***(0.008)
Percent Hispanic within male population		-0.020***(0.007)	-0.011**(0.006)	-0.015**(0.007)
Median household income (in 1000 s)		-0.111***(0.008)	-0.090***(0.005)	-0.101***(0.008)
Percent with access to exercise space		-0.027***(0.003)		-0.025***(0.003)
Percent food insecure		-0.126***(0.039)		-0.120***(0.04)
Percent drive alone to work		0.132***(0.009)		0.132***(0.01)
Violent crime rate		0.0001(0.0004)		0.0001(0.0004)
Fast food restaurants (per 1000 residents)		-0.088***(0.019)		-0.071***(0.019)
Grocery stores (per 1000 residents)		-0.034(0.032)		-0.018(0.035)
Constant	32.301***(0.475)	33.919***(1.502)	38.007***(0.768)	32.755***(1.548)
Observations	2921	2909	2921	2767
R-Squared	0.579	0.597	0.598	0.609

Note: Coefficients and standard errors (in brackets) for four state-level linear mixed effects models for male obesity prevalence. The models incorporate Twitter variables, demographic variables, and built environment variables. From left to right: (1) a model that includes only Twitter variables, (2) a model that includes demographic and built environment variables, (3) a model that includes Twitter and demographic variables and (4) a model that includes Twitter, demographic and built environment variables. Similar to the models for females, we noted significant negative associations between exercise tweet volume and obesity, and significant negative associations between high food tweet volume and obesity, while controlling for demographic and built environment variables.
*p < 0.1; **p < 0.05; ***p < 0.01

Table 2 Coefficients and standard errors for linear mixed-effects models for female obesity prevalence

	Twitter variables	Built environment and demographic variables	Twitter and demographic variables	Twitter, demographic and built environment variables
Tweet sentiment toward food	-0.024***(0.005)		-0.012***(0.004)	-0.006(0.004)
Percent of tweets food related (2nd tertile)	0.011(0.193)		0.101(0.164)	0.267*(0.162)
Percent of tweets food related (3rd tertile)	-1.023***(0.232)		-0.693***(0.197)	-0.366*(0.196)
Percent of tweets fast food related (2nd tertile)	0.074(0.176)		0.18(0.153)	0.238(0.155)
Percent of tweets fast food related (3rd tertile)	-0.402**(0.194)		-0.233(0.173)	-0.131(0.178)
Percent of tweets healthy food related (2nd tertile)	-0.369**(0.185)		-0.237(0.158)	-0.106(0.156)
Percent of tweets healthy food related (3rd tertile)	-0.496**(0.210)		-0.276(0.179)	-0.118(0.177)
Tweet sentiment toward exercise	-0.010**(0.004)		-0.005(0.003)	-0.003(0.003)
Percent of tweets exercise related (2nd tertile)	-1.095***(0.177)		-0.386**(0.152)	-0.367**(0.149)
Percent of tweets exercise related (3rd tertile)	-1.782***(0.186)		-0.766***(0.163)	-0.679***(0.162)
Median age within female population		-0.120***(0.014)	-0.084***(0.014)	-0.101***(0.015)
Percent non-Hispanic black within female population		0.136***(0.009)	0.127***(0.006)	0.133***(0.009)
Percent Hispanic within female population		-0.006(0.008)	-0.017**(0.007)	-0.006(0.009)
Median household income (in 1000 s)		-0.115***(0.009)	-0.131***(-0.006)	-0.115***(-0.01)
Percent with access to exercise space		-0.030***(0.003)		-0.029***(0.003)
Percent food insecure		0.012(0.045)		-0.005(0.046)
Percent drive alone to work		0.124***(0.011)		0.127***(0.012)
Violent crime rate		0.0003(0.0004)		0.0004(0.0004)
Fast food restaurants (per 1000 residents)		-0.110***(0.022)		-0.099***(0.023)
Grocery stores (per 1000 residents)		-0.045(0.037)		-0.07(0.045)
Constant	32.154***(0.615)	31.793***(1.731)	39.792***(0.866)	31.666***(1.795)
Observations	2876	2909	2876	2733
R-Squared	0.564	0.669	0.647	0.669

Note: Coefficients and standard errors (in brackets) for four state-level linear mixed effects models for female obesity prevalence. The models incorporate Twitter variables, demographic variables, and built environment variables. From left to right: (1) a model that includes only Twitter variables, (2) a model that includes demographic and built environment variables, (3) a model that includes Twitter and demographic variables and (4) a model that includes Twitter, demographic and built environment variables. We consistently observed significant negative associations between exercise tweet volume and obesity, and significant negative associations between high food tweet volume and obesity, while controlling for demographic and built environment variables.
*p < 0.1; **p < 0.05; ***p < 0.01

specific changes in behavioral risk factors for obesity at the US county level.

We observe several similarities between the data for males and females. However, major disparities such as, differences in the mention of high intensity physical activities (Cesare et al., 2019) agrees with data suggesting that women are less likely to meet physical activity recommendations compared to men (Centers for

Disease Control and Prevention, 2018b). Furthermore, women report foods with higher calories and have an overall higher obesity prevalence (Ogden et al., 2014; Flegal et al., 2016; Tauqeer et al., 2018). These observations support the need for separate approaches for estimating male and female obesity prevalence. In addition, although the Twitter population is not completely representative of the census demographic distributions, our

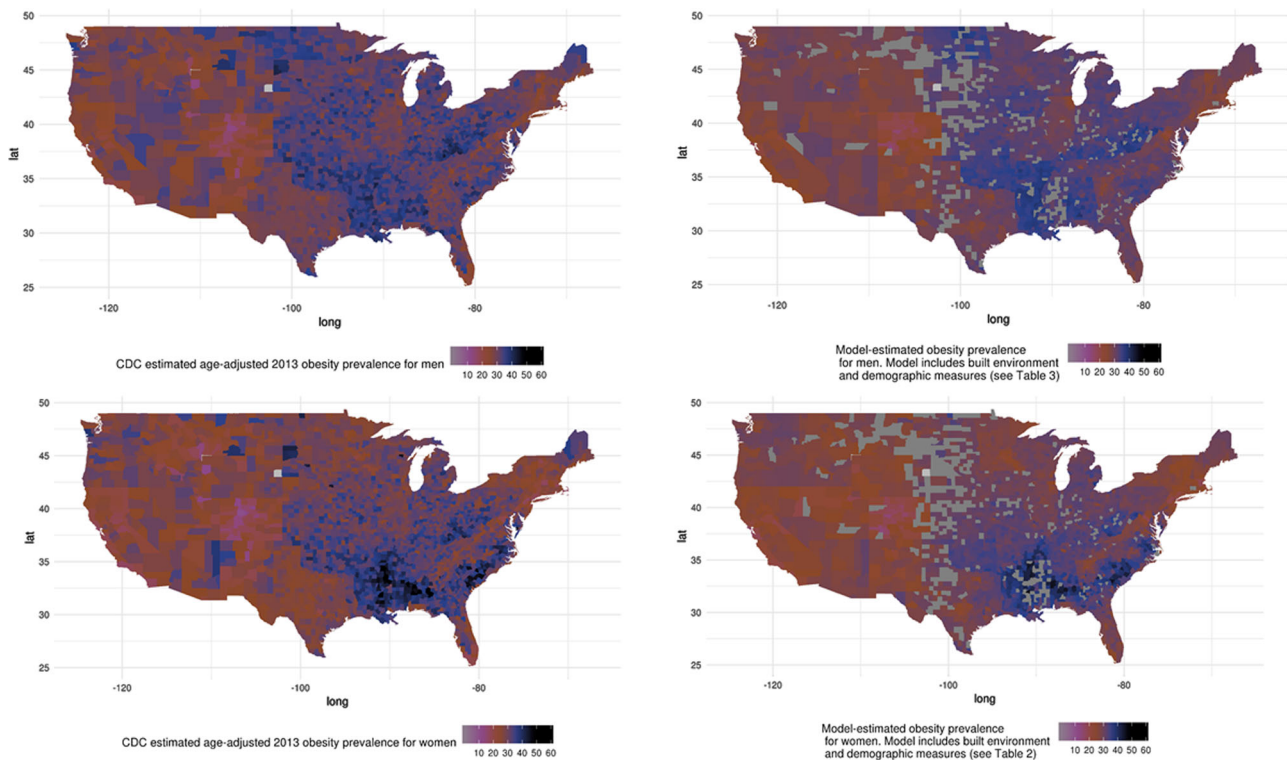


Fig. 2 Estimates of obesity prevalence. We estimated obesity prevalence separately for males and females using the previously described mixed-effects models with five-fold cross validation. These plots suggest that our models could provide good estimates of spatial trends in obesity prevalence. The gray areas indicate missing data

findings can benefit obesity prevention and intervention programs that target individuals who fall within the age group of Twitter users.

One factor that could have affected the outcomes in this study is a lack of survey data that overlaps with the time period of the Twitter data. Obesity estimates at the county level were produced by the CDC using 2013 BRFSS data, while our food and exercise tweets were collected from 2015 through 2016. However, we do not anticipate that this temporal difference significantly impacts our results. As previously noted, data from the State of Obesity Project suggest the increase in obesity prevalence between 2013 and 2016 was only significant for three states: Alabama, Michigan and Nebraska (see SI Fig. 1). Also, to accommodate this difference we forecasted 2015 sex-specific county-level obesity estimates using an autoregressive linear model that incorporates data from 2004 to 2013 and found that our model effects remained consistent (see SI Table 2). The effect direction and significance remain the same across covariates, and coefficients are comparable across models.

Another limitation is the use of median household income to estimate county-level SES. Studies suggest that aggregated income is not a comprehensive measure of SES status, particularly as it relates to health outcomes (Duncan et al., 2002; Pardo-Crespo et al., 2013). Factors such as personal wealth (Duncan et al., 2002) or local cost of living (Broda et al., 2009) may also impact health habits by influencing access to exercise spaces and healthy food. Future research will explore the inclusion of county-level SES estimates that extends beyond median income.

Despite these limitations, the findings in this study are promising for future research and applications that take advantage of the massive amounts of data generated daily on digital platforms to improve population health. Obesity is a complex health issue that has been associated with poverty in the United States among other factors (Zhang and Wang, 2004; Ogden et al., 2010).

Individuals in lower income neighborhoods tend to lack access to resources that encourage healthy behaviors (such as, grocery stores with healthy food options and safe spaces for physical activity). By developing methods that allow for near real-time monitoring of health behaviors at different scales, we can produce timely community level indicators of health that would enable timely comparisons across various demographic and socioeconomic factors. Internet data should be combined with other data streams including, sociodemographic, built environment variables and data from mobile applications and wearables, to improve the speed at which researchers and policymakers can monitor small area changes in sex-specific obesity.

Data availability

All data analyzed are included in the published paper.

Received: 23 April 2019 Accepted: 13 August 2019

Published online: 17 September 2019

References

- Ainsworth BE, Haskell WL, Whitt MC, Irwin ML, Swartz AM, Strath SJ, O'Brien WL, Bassett DR, Schmitz KH, Emplaincourt PO (2000) Compendium of physical activities: an update of activity codes and MET intensities. *Med Sci Sports Exerc* 32(9 Suppl):S498–S504
- Bennett GG, Wolin KY, James SA (2007) Lifecourse socioeconomic position and weight change among blacks: the Pitt County study. *Obesity* 15(1):172–172. <https://doi.org/10.1038/oby.2007.522>
- Broda C, Leibtag E, Weinstein DE (2009) The role of prices in measuring the poor's living standards. *J Econ Perspect* 23(2):77–97. <https://doi.org/10.1257/jep.23.2.77>
- Burger JD, Henderson J, Kim G, Zarrella G (2011) Discriminating gender on Twitter. In: Proceedings of the Conference on empirical methods in natural language processing. Association for Computational Linguistics, Stroudsburg, pp 1301–1309

- Casagrande SS, Whitt-Glover MC, Lancaster KJ, Odoms-Young AM, Gary TL (2009) Built environment and health behaviors among African Americans: a systematic review. *Am J Prev Med* 36(2):174–181
- Centers for Disease Control and Prevention (2018a) Behavioral risk factor surveillance system. <https://www.cdc.gov/brfss/index.html>. Accessed 12 July 2018
- Centers for Disease Control and Prevention (2018b) Adult Obesity Facts | Overweight and Obesity. Centers for disease control and prevention. <https://www.cdc.gov/obesity/data/adult.html>. Accessed Mar 21 2018
- Cesare N, Grant C, Hawkins JB, Brownstein JS, Nsoesie EO (2017a) Demographics in social media data for public health research: does it matter? Bloomberg Data for Good Exchange Conference, New York
- Cesare N, Grant C, Nsoesie EO (2017b) Detection of user demographics on social media: a review of methods and recommendations for best practices. Preprint at arXiv:1702.01807. <https://arxiv.org/abs/1702.01807>
- Cesare N, Nguyen QC, Grant C, Nsoesie EO (2019) Social media captures demographic and regional physical activity. *BMJ Open Sport Exercise Med* 5 (1). <https://doi.org/10.1136/bmjsem-2019-000567>
- Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. *New Engl J Med* 357(4):370–379. <https://doi.org/10.1056/NEJMsa066082>
- Chunara R, Bouton L, Ayers JW, Brownstein JS (2013) Assessing the online social environment for surveillance of obesity prevalence. *PLoS ONE* 8(4):e61373. <https://doi.org/10.1371/journal.pone.0061373>
- Cooksey-Stowers K, Schwartz MB, Brownell KD (2017) Food swamps predict obesity rates better than food deserts in the United States. *Int J Environ Res Public Health* 14(11). <https://doi.org/10.3390/ijerph14111366>
- County Health Rankings and Roadmaps (2016) University of Wisconsin Population Health Institute. County health rankings: how healthy is your County? <http://www.countyhealthrankings.org/homepage>. Accessed 19 Nov 2018
- De Choudhury M, Sharma S and Kiciman (2016) Characterizing Dietary Choices, Nutrition, and Language in Food Deserts via Social Media. In Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. CSCW '16. ACM, New York, pp 1157–1170
- Duncan GJ, Daly MC, McDonough P, Williams DR (2002) Optimal indicators of socioeconomic status for health research. *Am J Public Health* 92 (7):1151–1157
- Dwyer-Lindgren L, Freedman G, Engell RE, Fleming TD, Lim SS, Murray CJL, Mokdad AH (2013) Prevalence of physical activity and obesity in US counties, 2001–2011: a road map for action. *Population Health Metrics* 11(7). <https://doi.org/10.1186/1478-7954-11-7>
- Efron B, Tibshirani R (1986) Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Stat Sci* 1(1):54–75
- Ellaway A, Macintyre S, Bonnefoy X (2005) Graffiti, greenery, and obesity in adults: secondary analysis of European cross sectional survey. *BMJ* 331:611–612
- Flegal KM, Kruszon-Moran D, Carroll MD, Fryar CD, Ogden CL (2016) Trends in obesity among adults in the United States, 2005 to 2014. *JAMA* 315 (21):2284–2291. <https://doi.org/10.1001/jama.2016.6458>
- Freedman DS, Khan LK, Serdula MK, Galuska DA, Dietz WH (2002) Trends and correlates of class 3 obesity in the United States from 1990 through 2000. *JAMA* 288(14):1758–1761
- Fryar CD, Carroll MD, Ogden CL (2016) Prevalence of Overweight, Obesity, and Extreme Obesity among Adults Aged 20 and Over: United States, 1960–1962 Through 2013–2014. National Center for Health Statistics: Health E-Stats. https://www.cdc.gov/nchs/data/hestat/obesity_adult_13_14/obesity_adult_13_14.pdf. Accessed 13 Sep 2017
- Giles-Corti B, Macintyre S, Clarkon JP, Pikora T, Donovan RJ (2003) Environmental and lifestyle factors associated with overweight and obesity in Perth, Australia. *Am J Health Promot* 18(1):93–102
- Gordon-Larsen P, Nelson MC, Page P, Popkin BM (2006) Inequality in the built environment underlies key health disparities in physical activity and obesity. *Pediatrics* 117(2):417–424. <https://doi.org/10.1542/peds.2005-0058>
- Griffin GP, Jiao J (2015) Where does bicycling for health happen? Analysing volunteered geographic information through place and plexus. *J Transp Health* 2(2):238–247. <https://doi.org/10.1016/j.jth.2014.12.001>
- Harvard Health Publications (2015) Calories burned in 30min for people of three different weights. <https://www.health.harvard.edu/diet-and-weight-loss/calories-burned-in-30-minutes-of-leisure-and-routine-activities>. Accessed 26 June 2018
- Hill JO, Peters JC (1998) Environmental contributions to the obesity epidemic. *Science* 280(5368):1371–1374. <https://doi.org/10.1126/science.280.5368.1371>
- Hill JO, Wyatt HR, Reed GW, Peters JC (2003) Obesity and the environment: where do we go from here? *Science* 299(5608):853–855. <https://doi.org/10.1126/science.1079857>
- Jestic B, Nelson T, Winters M (2016) Mapping ridership using crowdsourced cycling data. *J Transp Geogr* 52:90–97. <https://doi.org/10.1016/j.jtrangeo.2016.03.006>
- Joulin A, Grave E, Piotr B, Tomas M (2016) Bag of tricks for efficient text classification. Preprint at arXiv:1607.01759 [cs]. Accessed 4 Feb 2019
- Kaggle. Sentiment classification (2011) <https://inclass.kaggle.com/c/si650winter11>. Accessed 16 Aug 2016
- Longley PA, Adnan M, Lansley G (2015) The geotemporal demographics of Twitter usage. *Environ Plan A* 47(2):465–484. <https://doi.org/10.1068/a130122p>
- Lopez-Zetina J, Lee H, Friis R (2006) The link between obesity and the built environment. Evidence from an ecological analysis of obesity and vehicle miles of travel in California. *Health Place* 12(4):656–664
- Maharana A, Nsoesie EO (2018) Use of deep learning to examine the association of the built environment with prevalence of neighborhood adult obesity. *JAMA Netw Open* 1:e181535–e181535
- Malec D, Sedransk J, Moriarity CL, LeClere FB (1997) Small area inference for binary variables in the national health interview survey. *J Am Stat Assoc* 92 (439):815–826. <https://doi.org/10.2307/2965546>
- Mattes R, Foster GD (2014) Food environment and obesity. *Obesity* 22 (12):2459–2461. <https://doi.org/10.1002/oby.20922>
- Maurer D, Pathman T, Mondloch CJ (2006) The shape of boubas: sound-shape correspondences in toddlers and adults. *Dev Sci* 9(3):316–322. <https://doi.org/10.1111/j.1467-7687.2006.00495.x>
- McCallum A (2002) MALLET: A machine learning for language toolkit. <http://mallet.cs.umass.edu>. Accessed 27 Feb 2019
- McFerran B, Dahl DW, Fitzsimons GJ, Morales AC (2009) I'll have what she's having: effects of social influence and body type on the food choices of others. *J Consum Res* 36(6):915–929. <https://doi.org/10.1086/644611>
- Mislove A, Lehmann S, Ahn YY, Onnela JP, Rosenquist JN (2011) Understanding the Demographics of Twitter Users. In: Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media. AAAI Publications, Menlo Park, pp 554–557
- Mobley LR, Root ED, Finkelstein EA, Khavjou O, Farris RP, Will JC (2006) Environment, obesity, and cardiovascular disease risk in low-income women. *Am J Prev Med* 30(4):327–332
- Mueller J and Stumme G (2016) Gender Inference using Statistical Name Characteristics in Twitter. In: Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on Social Informatics. ACM Press, Albany, 1–8
- Nelson MC, Gordon-Larsen P, Song Y, Popkin BM (2006) Built and Social Environments: Associations with Adolescent Overweight and Activity. *Am J Prev Med* 31(2):109–117. <https://doi.org/10.1016/j.amepre.2006.03.026>
- Nesbit KC, Kolobe TH, Sisson SB, Ghement IR (2014) A model of environmental correlates of adolescent obesity in the United States. *J Adolesc Health* 55 (3):394–401. <https://doi.org/10.1016/j.jadohealth.2014.02.022>
- Nguyen QC, Li D, Meng H, Kath S, Nsoesie EO, Li F, Wen M (2016) Building a national neighborhood dataset from geotagged Twitter data for indicators of happiness, diet, and physical activity. *JMIR Public Health Surveill* 17(2):e158. <https://doi.org/10.2196/publichealth.5869>. PMC5088343. PMC
- Nguyen QC, McCullough M, Meng HW, Paul D, Li D, Kath S, Loomis G, Nsoesie EO, Wen M, Smith KR, Li F (2017) Geotagged US Tweets as Predictors of County-Level Health Outcomes, 2015–2016. *Am J Public Health* 107 (11):1776–1782. <https://doi.org/10.2105/AJPH.2017.303993>
- Nielsen A, Rendall D (2011) The sound of round: evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Can J Exp Psychol* 65(2):115–124. <https://doi.org/10.1037/a0022268>
- Nsoesie EO, Buckeridge LD, Brownstein JS (2014) Guess who's not coming to dinner? evaluating online restaurant reservations for disease surveillance. *J Med Internet Res* 16(1):e22. <https://doi.org/10.2196/jmir.2998>
- Nsoesie EO, Butler P, Ramakrishnan N, Mekaru SR, Brownstein JS (2015) Monitoring disease trends using hospital traffic data from high resolution satellite imagery: a feasibility study. *Sci Rep* 5:9112. <https://doi.org/10.1038/srep09112>. PMC4357853. PMC
- Ogden CL, Carroll MD, Kit BK, Flegal KM (2014) Prevalence of childhood and adult obesity in the United States, 2011–2012. *JAMA* 311(8):806–814. <https://doi.org/10.1001/jama.2014.732>
- Ogden CL, Lamb MM, Carroll MD and Flegal KM (2010) Obesity and Socio-economic Status in Children and Adolescents: United States, 2005–2008. National Center for Health Statistics: NCHSDData Brief 51, 1–8. <https://www.cdc.gov/nchs/data/databriefs/db51.pdf>. Accessed 5 Apr 2018
- Olson DR, Konty KJ, Paladini M, Viboud C, Simonsen L (2013) Reassessing Google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales. *PLoS Comput Biol* 9(10):e1003256. <https://doi.org/10.1371/journal.pcbi.1003256>
- Papas MA, Alberg AJ, Ewing R, Helzlsouer KJ, Gary TL, Klassen AC (2007) Built environment and obesity. *Epidemiol Rev* 29:129–143. <https://doi.org/10.1093/epirev/mxm009>
- Pardo-Crespo MR, Narla NP, Williams AR, Beebe TJ, Sloan J, Yawn BP, Wheeler PH, Juhn YJ (2013) Comparison of individual-level versus area-level socioeconomic measures in assessing health outcomes of children in Olmsted County, Minnesota. *J Epidemiol Community Health* 67(4):305–310. <https://doi.org/10.1136/jech-2012-201742>

- R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.gbif.org/en/tool/81287/r-a-language-and-environment-for-statistical-computing>
- Said A, Bellogin A (2014) You are What You Eat! Tracking Health Through Recipe Interactions. In: Proceedings of the 6th Workshop on Recommender Systems and the Social Web (RSWeb 2014) Foster City. https://pdfs.semanticscholar.org/6b9c/a6296deda297063f104bad16e4e2586301f4.pdf?_ga=2.36914276.572343149.1565202826-548947087.1565202826. Accessed 4 Apr 2018
- Sanders Analytics (2011) Twitter sentiment corpus. <http://www.sananalytics.com/lab/twitter-sentiment/>. Accessed 16 Aug 2016
- Segal LM, Rayburn J, Beck SE (2017) The State of Obesity: Better Policies for a Healthier America. The State of Obesity Project: Trust for America's Health and the Robert Wood Johnson Foundation. <https://www.stateofobesity.org/>. Accessed 29 Sep 2017
- Sentiment140 (2009) Sentiment 140: For Academics. <http://help.sentiment140.com/for-students> Accessed 16 Aug 2016
- Tauqeer Z, Gomez G, Stanford FC (2018) Obesity in women: insights for the clinician. *J Women's Health* (2002) 27(4):444–457. <https://doi.org/10.1089/jwh.2016.6196>
- Torres J, Ortiz K, García J, Vaca C (2018) Uncovering Aspects of Places for Fitness Activities Through Social Media. In: Proceedings of WorldCIST'18: Trends and Advances in Information Systems and Technologies. Advances in Intelligent Systems and Computing. Springer, Cham, pp 961–968
- Twitter Developers (2014) Difference between sample and filter streaming API. <https://twittercommunity.com/t/difference-between-sample-and-filter-streaming-api/15094>. Accessed 29 Sep 2016
- United States Census Bureau (2015) US Census Bureau's American Community Survey. <https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml>. Accessed 11 Jan 2019
- United States Department of Agriculture (2014) National Nutrient Database. <http://ndb.nal.usda.gov/ndb/search/list?format=&count=&max=25&sort=&fg=&man=&facet=&qlookup=&offset=50>. Accessed 28 Sep 2016
- United States Department of Agriculture (2018) Food environment Atlas. <https://www.ers.usda.gov/data-products/food-environment-atlas/data-access-and-documentation-downloads/>. Accessed 24 May 2018
- Wolpert DH (1992) Stacked generalization. *Neural Netw* 5(2):241–259
- Yakusheva O, Kapinos K, Weiss M (2011) Peer effects and the Freshman 15: evidence from a natural experiment. *Econ Hum Biol* 9(2):119–132. <https://doi.org/10.1016/j.ehb.2010.12.002>
- Yuan Q, Nsoesie EO, Lv B, Peng G, Chunara R, Brownstein JS (2013) Monitoring influenza epidemics in china with search query from baidu. *PLoS One* 8(5): e64323. <https://doi.org/10.1371/journal.pone.0064323>. 23750192
- Zhang N, Campo S, Janz KF, Eckler P, Yang J, Snetselaar LG, Signorini A (2013) Electronic word of mouth on twitter about physical activity in the United States: exploratory infodemiology study. *J Med Internet Res* 15(11):e261. <https://doi.org/10.2196/jmir.2870>
- Zhang Q, Wang Y (2004) Socioeconomic inequality of obesity in the United States: do gender, age, and ethnicity matter? *Soc Sci Med* 58(6):1171–1180. [https://doi.org/10.1016/s0277-9536\(03\)00288-0](https://doi.org/10.1016/s0277-9536(03)00288-0)

Acknowledgements

Nina Cesare and Elaine O. Nsoesie are supported by a grant (#73362) from the Robert Wood Johnson Foundation. Quynh C. Nguyen is supported by NIH grant 5K01ES025433.

Additional information

Supplementary information: The online version of this article (<https://doi.org/10.1057/s41599-019-0314-x>) contains supplementary material, which is available to authorized users.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019