BMJ Open Sport & Exercise Medicine

Social media captures demographic and regional physical activity

Nina Cesare,[1,2] Quynh C Nguyen,[3] Christan Grant,[4] Elaine O Nsoesie[1,2]

To cite: Cesare N, Nguyen QC, Grant C, *et al*. Social media captures demographic and regional physical activity. *BMJ Open Sport & Exercise Medicine* 2019;**5**:e000567. doi:10.1136/bmjsem-2019-000567

► Additional material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136bmjsem-2019-000567).

Accepted 2 July 2019

[1]Department of Global Health, Boston University School of Public Health, Boston, Massachusetts, USA
[2]Institute for Health Metrics and Evaluation, University of Washington, Seattle, Washington, USA
[3]Department of Epidemiology and Biostatistics, University of Maryland School of Public Health, College Park, Maryland, USA
[4]School of Computer Science, University of Oklahoma, Norman, Oklahoma, USA

Correspondence to
Dr Nina Cesare;
ncesare@bu.edu

ABSTRACT

Objectives We examined the use of data from social media for surveillance of physical activity prevalence in the USA.

Methods We obtained data from the social media site Twitter from April 2015 to March 2016. The data consisted of 1 382 284 geotagged physical activity tweets from 481 146 users (55.7% men and 44.3% women) in more than 2900 counties. We applied machine learning and statistical modelling to demonstrate sex and regional variations in preferred exercises, and assessed the association between reports of physical activity on Twitter and population-level inactivity prevalence from the US Centers for Disease Control and Prevention.

Results The association between physical inactivity tweet patterns and physical activity prevalence varied by sex and region. Walking was the most popular physical activity for both men and women across all regions (15.94% (95% CI 15.85% to 16.02%) and 18.74% (95% CI 18.64% to 18.88%) of tweets, respectively). Men and women mentioned performing gym-based activities at approximately the same rates (4.68% (95% CI 4.63% to 4.72%) and 4.13% (95% CI 4.08% to 4.18%) of tweets, respectively). CrossFit was most popular among men (14.91% (95% CI 14.52% to 15.31%)) among gym-based tweets, whereas yoga was most popular among women (26.66% (95% CI 26.03% to 27.19%)). Men mentioned engaging in higher intensity activities than women. Overall, counties with higher physical activity tweets also had lower leisure-time physical inactivity prevalence for both sexes.

Conclusions The regional-specific and sex-specific activity patterns captured on Twitter may allow public health officials to identify changes in health behaviours at small geographical scales and to design interventions best suited for specific populations.

What are the findings?

- Men mentioned engaging in higher intensity physical activities than women, which agrees with previous studies suggesting that women are less likely to meet recommendations for aerobic physical activity.
- There were differences in the types of physical activities reported across the four US regions.

How might it impact clinical practice?

- Differences in the types of physical activities reported across sex and regions in the US can encourage discussions between clinicians and patients regarding exercise choices for weight loss and cardiovascular health.
- In the future, with patient consent, clinicians can use individual patient reports on physical activity posted on social media for personalized guidance.

INTRODUCTION

Insufficient physical activity is considered a modifiable risk factor for non-communicable diseases (such as cardiovascular diseases and diabetes) and has been associated with loss of life[1 2] and significant global economic cost. In 2016, worldwide healthcare costs associated with physical inactivity were approximately $53.8 billion, and inactivity contributed to productivity losses of about $13.7 billion.[3] The WHO member states have agreed to develop and implement policies aimed at reducing physical inactivity rates by 10% by 2025.[4 5] Achieving this target requires timely surveillance of physical activity prevalence across populations.

In order to decrease physical inactivity prevalence, it is important to target interventions towards at-risk groups and regions with higher prevalence. However, estimating inactivity prevalence using traditional survey approaches can be costly and delayed, and may be subject to social desirability or recall bias.[6–8] Moreover, estimates of inactivity prevalence may not include information regarding which forms of exercise individuals prefer.

Digital technology, including sensors found in cell phones and wrist bands[9–11] and mobile fitness applications (such as RunKeeper or Strava), may be used to document physical activity.[12–14] While these tools provide valuable information about movement, social media applications (such as Twitter and Instagram) can provide insight into both preferred activities and attitudes towards physical activity.[15 16] Reports of physical activity on these platforms are not restricted to preset choices, thereby enabling the use of descriptive textual information that may publicly capture a diverse array of preferred activities, exercise intensity

and attitudes towards physical activity in real time and at scale. Timely estimates of physical activity prevalence from combining social media data with other data sources could be useful for monitoring spatial and temporal trends and for augmenting traditional methods for monitoring physical activity.

Here, we use data from Twitter to assess sex-specific similarities and differences in self-reported leisure-time physical activity across US counties and regions. First, we assess differences in the types of activities in which users report engaging and the intensity of these activities as measured by calories burned in 30 min of activity. Second, we quantify the association between estimates of physical inactivity from the US Centers for Disease Control and Prevention (CDC) and physical activity postings and sentiment while controlling for internet search activity, county demographics and environmental variables associated with physical activity.

METHODS

Extraction of Twitter data

Mentions of physical activity were retrieved from the Twitter streaming application programming interface (API) from April 2015 to March 2016 using a set of 376 keywords (online supplementary table S1) gathered from fitness questionnaires and apps that document a range of activities.[17 18] These included team sports, gym exercises, outdoor recreational activities and household chores. These data were collected as part of a larger project designed to assess the association between happiness and health indicators constructed from social media data with known public health measures. Initial findings using these data are reported in cited references.[19 20]

Data processing

Several steps were taken to clean and ensure data reliability. Content from users at the top 99th percentile of tweet activity was removed. A keyword-matching algorithm was used to identify relevant tweet content. Tweets containing popular phrases that denote irrelevant content (eg, 'walk away' or 'running late') and mentions of the television show 'Walking Dead' were removed. For team sports, only tweets that contained the word play/playing/played in conjunction with the activity were retained. This was to ensure that only tweets indicating engagement in physical activity were used in our analysis. To assess the performance of this algorithm, a subset of categorised tweets was compared with hand-labelled tweets. The accuracy was 85% with an F1 score (defined as the harmonic mean of the precision and recall/sensitivity; 1 is the highest value) of 0.90. We also tested several supervised machine learning classifiers, including a feed-forward neural network, support vector machines (SVMs), gradient boosting and fastText.[21] The keyword matching algorithm performed best.

Exercise intensity was calculated as calories burned during 30 min of a particular physical activity by a 155 lb individual, the average weight of an American adult.[17 22] Each tweet was mapped to a US county based on its geocode (ie, latitude and longitude). See online supplementary table S2 for a sample of tweets. Note that select words and characters in these tweets (online supplementary table S2) have been changed to maintain user privacy (eg, 'went running with the bf' may be changed to 'went running with my bf').[23]

Analysing sentiment

A maximum entropy text classifier in Java's MALLET toolkit and ground truth data from Kaggle, Sentiment 140 and Sanders Analytics were used to categorise tweet sentiment and to assign each tweet a 'happiness' score between 0 and 1. As previously noted, this dataset was originally constructed to assess the association between happiness and health indicators constructed from social media data with public health outcomes. Tweets with a score of 0.80 or higher were classified as 'happy'. Sentiment, in the context of this study represents the proportion of tweets that are labelled as happy. For more information on data processing and sentiment analysis, refer to cited references.[19 20]

Classification of Twitter users' sex

We applied a previously developed gender classifier that employs a weighted ensemble approach and uses features/information from users' names to predict whether users are male or female. We acknowledge that there is a distinction between gender and sex (West and Zimmerman, 1987), but we use gender estimates as a proxy for sex in order to be consistent with CDC measures. This technique combines three classification approaches: (a) matching users' first names to data from the US Social Security Administration[24] (which captures approximately 60% of Twitter names), (b) an SVM classifier applied to word and character n-gram features from users' names[25] and (c) a decision tree classifier applied to features constructed from the linguistic structure of users' names, including the count of syllables, vowels, consonants, bouba (round) and kiki (sharp) vowels and consonants,[26 27] and whether or not the last character is a vowel.[28] For each user, we combined the predictions from all three classifiers using a weighted stacked logistic regression framework.[29] The ensemble classifier achieved an accuracy of 0.83, a recall of 0.85 and an F1 score of 0.84. The ensemble classifier performed better than methods b and c, and captured all users with alphanumeric names, unlike method a. See these papers by Cesare *et al.* for a detailed description of this classification framework.[30 31]

Leisure-time physical inactivity estimates

We obtained 2009–2013 county-level measures of leisure-time physical inactivity from the US CDC. These measures were generated from self-reported physical activity engagement from the Behavioural Risk Factor and Surveillance System survey using small-area estimation techniques.[32] Note that the most recent estimates

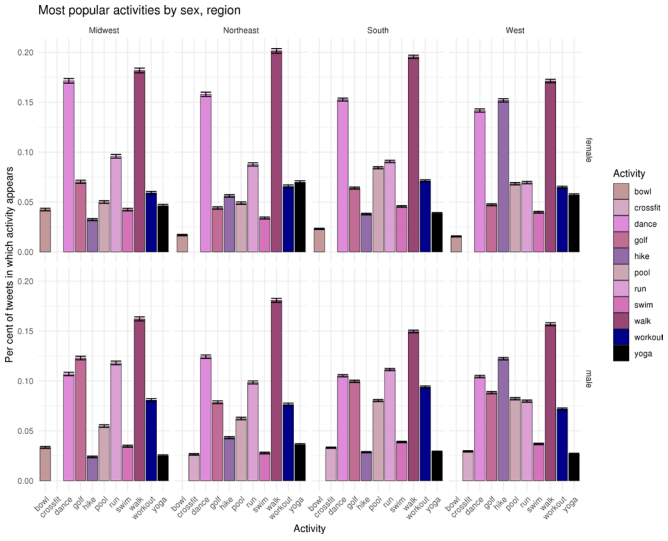

Figure 1 The 10 most frequently mentioned activities and the proportion of tweets represented by sex and region.

from 2013 do not overlap with our Twitter data, which wre collected between 2015 and 2016. We therefore used linear autoregressive models to forecast physical inactivity prevalence 2 years ahead. When predicting 2013 inactivity prevalence based on 2011 data, our models captured approximately 88% and 83% variations in inactivity estimates for men and women, respectively. We applied the same models to forecast physical inactivity for 2015.

Google trends

Google Trends provides an index between 1 and 100 representing relative search volume related to terms and topics across time and space. We searched for terms related to physical activity and fitness fads, including 'intermittent fasting', 'workout', 'fitness centre', 'gym', 'weight loss' and 'physical fitness' within the same time span as our Twitter data (April 2015–March 2016). To select which terms may be relevant in our model, we analysed the correlation between each term and inactivity prevalence, as well as the correlation between each term and the others to avoid multicollinearity. We selected the terms fitness centre and weight loss for this analysis. The geographical distribution of Google Search Index values used can be found in online supplementary figure S1.

Statistical analysis

To examine the association between indicators of physical activity constructed from Twitter data and county-level estimates of physical inactivity based on CDC data, we used a series of linear mixed-effects regression models with varying state intercepts.[33]

We accounted for socioeconomic and demographic variables that have been associated with inactivity prevalence.[34–39] These included median household income, racial/ethnic composition and median county age. Per cent non-Hispanic white was strongly correlated with per cent non-Hispanic black and per cent Hispanic. We therefore used only per cent non-Hispanic black and per cent Hispanic. These data were obtained from the 2015 5-year American Community Survey.

We also account for environmental factors that may impact community health, including community and road safety and access to usable exercise space.[40–42] We obtained data on the per cent of individuals in each county who have access to exercise space, the violent crime rate in the county and the rate of driving deaths (measured as the total number of driving deaths divided by the county population) from the County Health Rankings and Roadmap project.[43] Data on walkability were unavailable at the county level. We report our findings with 95% CI.

Patient and public involvement

There were no patients involved in this study. We used publicly available data, but members of the public were not involved to comment on study design, to interpret results or to contribute to the writing and editing of this document.

RESULTS

There were sex and regional differences in physical activities

Our analysis included 1 382 284 physical activity geotagged tweets (80 million tweets collected) from 481 146 users (55.65% men and 44.35% women) in 2992 and 2932 counties, respectively, for men and women. We grouped our findings into four geographical regions in the USA: West, South, Northeast and Midwest.[44] Sentiment towards exercise was distributed with some uniformity across the USA (online supplementary figure S2). Overall, men and women shared similar sentiments towards physical activity (sentiment scores 0.660 (95% CI 0.660 to 0.661) and 0.657 (95% CI 0.656 to 0.657), respectively).

The top exercise terms were 'walk', 'dance', 'golf', 'workout', 'run', 'pool', 'hike', 'yoga', 'swim' and 'bowl'. Walking was the most popular physical activity for both groups across all regions. We note overall and sex-specific regional variations in preferred activities (figure 1). For women, hiking was the second most popular activity in the West, representing 15.18% (95% CI 15.01% to 15.35%) of tweets. This activity represented only 3.24% (95% CI 3.13% to 3.35%) to 3.79% (95% CI 3.71% to 3.87%) of tweets in the Midwest and South, respectively. Mentions of participation in yoga also varied by region for women, representing 6.97% (95% CI 6.83% to 7.12%) of tweets in the Northeast, but only 3.87% (95% CI 3.79% to 3.95%) of tweets in the South. We saw similar patterns for hiking among men, representing 12.23% (95% CI 12.10% to 12.36%) of tweets in the West but only 2.38% (95% CI 2.30% to 2.47%) to 4.31% (95% CI 4.20% to 4.42%) of tweets elsewhere. Golf also varied in popularity among men, representing 12.29% (95% CI 12.11% to 12.48%) of tweets in the Midwest but only 7.85% (95% CI 7.72% to 8.00%) of tweets in the Northeast.

Men and women mentioned performing gym-based activities at approximately equivalent rates (4.68% (95% CI 4.63% to 4.72%) and 4.13% (95% CI 4.08% to 4.18%)

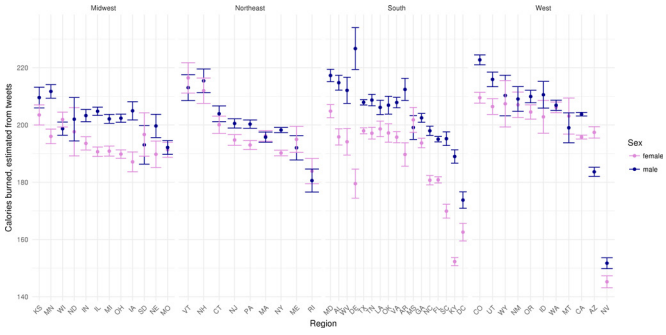

Figure 2 State-level comparisons of self-reported calories burned estimated based on physical activities mentioned by men and women on social media by state and region. Sex-based disparities are on average more significant in the South.

of tweets, respectively). CrossFit was the most popular workout class among men (14.91% (95% CI 14.52% to 15.31%) of gym-based tweets), whereas yoga was the most popular workout class among women (26.66% (95% CI 26.03% to 27.19%) of gym-based tweets). However, there were some differences, although not significant, in the estimated intensity of exercises reported by men and women as measured in calories burned (figure 2). The average number of reported calories burned per 30 min of reported exercise was 201.27 (95% CI 201.01 to 201.54) for men, and 191.66 (95% CI 191.37 to 191.95) for women. There were also regional variations in reported exercise intensity within sex. Women in the West reported exercises with the highest average caloric expenditure (ie, 194.78 (95% CI 194.25 to 195.31)), followed by the Northeast (193.26 (95% CI 192.60 to 193.92)), the Midwest (192.62 (95% CI 191.92 to 193.32)) and the South (187.96 (95% CI 187.48 to 188.44)). In contrast, men in the Midwest reported exercises with the highest caloric expenditure (202.71 (95% CI 202.07 to 203.36), followed by the South (202.58 (96% CI 202.13 to 203.04), the West (200.34 (95% CI 199.86 to 200.83)) and the

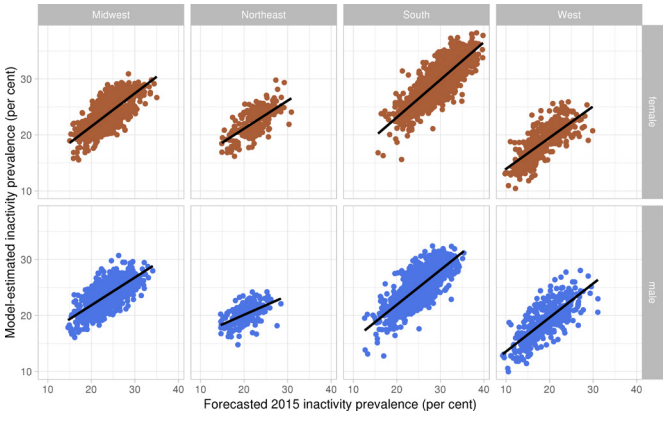

Figure 3 The relationship between model-estimated and Centers for Disease Control and Prevention-forecasted inactivity prevalence based on mixed-effects linear models that control for measures of physical activity via Twitter, demographic variables and built environment contextual variables. Lines represent a linear fit.

Northeast (198.93 (95% CI 198.32 to 199.54)). The most significant sex disparities were noted in counties within Southern states; the average difference between men and women was 8.51 calories per 30 min of activity.

Overall, counties that reported higher levels of physical activity on Twitter also had lower physical inactivity prevalence

The proportion of exercise tweets in a county was negatively associated with leisure-time physical inactivity prevalence for both men and women across regions (see figure 3). These correlations were strongest in the Northeast (r=−0.234 and −0.373 for men and women, respectively) and in the West (r=−0.217 and −0.267 for men and women, respectively). The national association between tweet sentiment and physical inactivity was similar for both men and women (r=−0.115 for men and r=−0.116 for women), but regional disparities exist. This relationship was stronger for men in the West (r=−0.194 for men and r=−0.076 for women) and stronger for women in the Northeast (r=−0.271 for women and r=−0.063 for men). There was a weak negative association between exercise intensity and physical inactivity for both groups (r=−0.061 for men and −0.001 for women), but stratified by region, this effect was much stronger for men in the West (r=−0.203) and the Midwest (r=−0.123). The association between each Twitter variable and inactivity prevalence by sex and region can be found in online supplementary table S3.

The association between Twitter variables and inactivity remained in models that accounted for demographic, socioeconomic and built environment variables associated with physical inactivity (tables 1 and 2). This relationship was statistically significant for all regions for men, and all regions except the Midwest for women. Also, counties with more positive sentiment towards physical activity had lower inactivity prevalence in the West for both men and women, and in the Midwest for women. Furthermore, counties that reported high-intensity exercises on Twitter also had lower inactivity prevalence for men in the Midwest and the Northeast. There was no significant relationship between exercise intensity and physical inactivity prevalence for women.

We also observed different patterns in the association between Google searches for fitness centres and weight loss and physical inactivity prevalence in the two demographic groups. Specifically, counties in the Northeast with higher searches for 'fitness centres' also had lower physical inactivity for women, while counties in the Northeast and South with higher searches for weight loss had higher inactivity for women. Among men, counties with higher searches for fitness centre had lower inactivity prevalence, while counties with higher weight loss searches had higher inactivity prevalence in the Midwest. The directionality of these relationships suggests that populations seeking weight loss information online tend to have higher physical inactivity prevalence, while those

Table 1 Mixed-effects regression for county-level, female-specific inactivity by region

	Midwest	Northeast	South	West
Percent of exercise-related tweets (logged)	−0.205	−0.813*	−0.421***	−0.725***
	0.129	0.421	0.117	0.193
Tweet sentiment towards exercise	−0.008*	−0.019	−0.005	−0.016**
	0.004	0.021	0.004	0.006
Average exercise intensity, via Tweets	0.001	0.01	0.002	0.0002
	0.002	0.007	0.002	0.002
'Fitness centre' Google search index	−0.068	−0.042*	0.051	0.015
	0.054	0.025	0.037	0.036
'Weight loss' Google search index	0.019	0.166**	0.136***	0.055
	0.078	0.068	0.044	0.078
Median age	−0.019	−0.065	−0.070***	−0.044*
	0.02	0.059	0.018	0.024
Median household income (in 1000s)	−0.094***	−0.060***	−0.105***	−0.044***
	0.01	0.018	0.007	0.013
Percent non-Hispanic black	0.011	0.032	0.024***	0.274***
	0.024	0.044	0.006	0.079
Percent Hispanic	0.041**	0.079***	−0.028***	0.008
	0.016	0.03	0.007	0.011
Percent with access to exercise space	−0.019***	−0.029**	−0.027***	−0.044***
	0.005	0.014	0.004	0.007
Driving death rate	0.850***	1.099**	0.912***	1.113***
	0.177	0.537	0.151	0.217
Violent crime rate	0.001	0.002	0.0003	0.002**
	0.001	0.002	0.0004	0.001
Constant	38.536***	27.252***	30.618***	29.222***
	5.267	7.346	4.201	5.796
Observations	888	214	1289	377
Adjusted R^2	0.620	0.560	0.698	0.629
SD, random intercept	2.404	1.075	1.817	2.033

Note: *p<0.1; **p<0.05; ***p<0.01.

seeking information on fitness centres are more likely to be active.

Notable disparities exist in the 2013 and 2015 (forecasted) prevalence of physical inactivity between men and women, with 79.7% and 78.4% of counties showing higher prevalence for women, respectively (figure 43). Our estimates of physical inactivity prevalence using physical activity tweet volume and sentiment, and Google search volume, while controlling for county demographics, and access to exercise space, are overall reflective of the disparities reported in CDC physical inactivity prevalence estimates. Overall, out-of-sample estimates are better for women than for men (average r=0.89 for women, average r=0.82 for men). Correlations between estimated and actual values were higher in the South for both men and women (r=0.79 and r=0.82, respectively).

DISCUSSION

This is the first study to assess the effectiveness of using Twitter for monitoring physical activity for men and women separately. Our main findings were (1) there were sex and regional differences in physical activities reported on Twitter, and (2) counties that reported higher levels of physical activity on Twitter tended to have lower physical inactivity prevalence based on survey estimates from the US CDC.

Some regions demonstrated higher associations between Twitter postings and survey estimates of physical inactivity from the CDC than others. For example, both tweet volume and sentiment were negatively associated with inactivity in the West for men and women. This agrees with research suggesting that the lowest inactivity prevalence in the USA is in the West.[37] Furthermore, the popularity of 'hiking'—an outdoor activity—in the West

Table 2 Mixed-effects regression for county-level, male-specific inactivity by region

	Midwest	Northeast	South	West
Percent of exercise-related tweets (logged)	−0.222*	−0.839**	−0.488***	−0.913***
	0.127	0.402	0.107	0.205
Tweet sentiment towards exercise	0.003	−0.011	0.001	−0.026***
	0.005	0.016	0.004	0.008
Average exercise intensity, via Tweets	−0.003**	0.010*	0.0003	−0.002
	0.002	0.006	0.001	0.003
'Fitness centre' Google search index	−0.090***	−0.014	0.043	0.017
	0.03	0.015	0.037	0.04
'Weight loss' Google search index	0.142***	0.055	0.063	−0.011
	−0.044	−0.036	−0.044	−0.088
Median age	0.035*	−0.002	−0.073***	−0.012
	0.021	0.059	0.018	0.025
Median household income (in 1000s)	−0.076***	−0.059***	−0.101***	−0.048***
	0.011	0.017	0.007	0.013
Percent non-Hispanic black	−0.026	0.032	−0.014**	0.216***
	0.026	0.041	0.006	0.081
Percent Hispanic	0.039**	0.074**	−0.026***	0.016
	0.017	0.029	0.007	0.012
Percent with access to exercise space	−0.027***	−0.038***	−0.028***	−0.037***
	0.005	0.014	0.004	0.007
Driving death rate	1.073***	1.128**	0.902***	1.131***
	0.188	0.502	0.145	0.221
Violent crime rate	−0.0002	0.0002	−0.0002	0.001
	0.001	0.002	0.0004	0.001
Constant	30.847***	29.785***	33.294***	32.998***
	3.421	5.734	4.121	6.458
Observations	909	213	1300	383
Adjusted R^2	0.527	0.380	0.634	0.616
SD, random intercept	1.332	0.319	1.791	2.299

Note: *p<0.1; **p<0.05; ***p<0.01.

comports with research that finds a negative association between urbanicity and inactivity in this region.[45] High levels of inactivity have been documented in the South, particularly among women.[46 47] Interestingly, we noted a strong negative relationship between tweet volume and inactivity, and a strong positive association between Google searches for weight loss and inactivity in this region. These may be important measures to monitor for addressing female-specific inactivity in this region.

Our findings also support research that states women are less likely to meet federal physical activity guidelines compared with men.[48] Women reported lower intensity exercises on Twitter compared with men, particularly in the South. Public health officials could focus on promoting other forms of leisure-time physical activity, such as transportation-based activity, which has documented health benefits.[49 50] Future analysis will focus on assessing deviations in activities undertaken by diverse sex and age populations during different times of the year, which may lead to more effective targeting of interventions.

Sex and regional deviations indicate that social media-based physical activity interventions should not be applied uniformly across the USA. Monitoring physical activity prevalence using social media and other digital sources can enable timely, geographically fine-grained estimates compared with surveys, thereby allowing for early intervention aimed at improving health over the life course. However, social media use varies by sex[51] and may also vary by place of residence, age or ethnicity. Public health officials should understand how individuals within specific groups use social media and target areas using appropriate platforms.

Findings from this study also suggest the need to combine data from Twitter with other digital sources because Twitter users may be more likely to report

physical activities that involve social interaction. Combining Twitter data with other data sources might mitigate limitations inherent in a single digital source.

LIMITATIONS

The measure of exercise intensity is arbitrary since actual reported exercise duration is unknown. This might have affected the reported similarities and differences in exercise intensity across sex and regions. There are also several forms of bias that may impact measures of physical activity from Twitter. For one, individuals have a tendency to over-report or overestimate their actual activity time and exertion.[52–54] Additionally, reports of physical activity on Twitter may be subject to selective residential bias. Individuals who are more likely to electively discuss personal physical activity may elect to live in areas that are more socially and structurally amenable to physical activity.[55–58] Furthermore, reports of physical activity on Twitter may not precisely correspond with engagement in physical activity; individuals may reflect on recent activity or express intent to engage in a physical activity. Finally, these data are subject to representation bias as well. Because our unit of analysis is the county, regions in which county density is proportionally lower are less likely to be represented. Given that several of the most densely populated states are in the Northeast, for instance (eg, Rhode Island, Massachusetts and Connecticut),[59] this region may be over-represented in our data.

One other limitation is the lack of county-level estimates of leisure-time physical inactivity from the CDC for 2015. We address this limitation by using autoregressive linear models to predict values for 2015 based on data from 2009 to 2013. These estimates are approximate, and interpretation of coefficients must consider this uncertainty. Future work will develop processes for combining non-traditional data sources to estimate small-area health outcomes, which are currently delayed by years.

CONCLUSIONS

Digital data, including social media, provide valuable information for monitoring health behaviours.[16 19 20 60–65] This study illustrates that Twitter is a useful tool for measuring small-area trends in physical activity, an important risk factor for non-communicable diseases, but that its usefulness might vary by sex and by US region. Monitoring physical activity using social media will allow public health officials to identify changes in health behaviours at small geographical scales across the USA. Findings from this study provide an important step in this direction.

Acknowledgements None

Contributors EON designed the study. QCN provided data and helped guide the analyses. CG guided data management and processing. NC and EON conducted the analyses, and wrote the initial draft of the paper. All authors edited the paper.

Funding NC, CG and EON are supported by a grant (#73362) from the Robert Wood Johnson Foundation. QCN is supported by National Institutes of Health grant 5K01ES025433.

Competing interests None.

Patient consent for publication Not required.

Ethics approval The study was declared exempt by the University of Washington Institutional Review Board.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon reasonable request.

REFERENCES

1. World Health Organization. Physical Inactivity: A Global Public Health Problem [Internet]. Global Strategy on Diet, Physical Activity and Health, 2019. Available: https://www.who.int/dietphysicalactivity/factsheet_inactivity/en/ [Accessed cited 14 Jan 2019].
2. Lim SS, Vos T, Flaxman AD, *et al*. A comparative risk assessment of burden of disease and injury attributable to 67 risk factors and risk factor clusters in 21 regions, 1990–2010: a systematic analysis for the global burden of disease study 2010. *The Lancet* 2012;380:2224–60.
3. Ding D, Lawson KD, Kolbe-Alexander TL, *et al*. The economic burden of physical inactivity: a global analysis of major non-communicable diseases. *The Lancet* 2016;388:1311–24.
4. Kohl HW, Craig CL, Lambert EV, *et al*. The pandemic of physical inactivity: global action for public health. *The Lancet* 2012;380:294–305.
5. Das P, Horton R. Physical activity—time to take it seriously and regularly. *The Lancet* 2016;388:1254–5.
6. van de Mortel TF. Faking it: social desirability response bias in self-report research. *Aust J Adv Nurs* 2005;25:40–8.
7. Althubaiti A. Information bias in health research: definition, pitfalls, and adjustment methods. *J Multidiscip Healthc* 2016;9:211–7.
8. Hassan E. Recall bias can be a threat to retrospective and prospective research designs. *Int J Epidemiol* 2005;3.
9. Song L, Mercer L, Wakefield J, *et al*. Using small-area estimation to calculate the prevalence of smoking by Subcounty geographic areas in King County, Washington, behavioral risk factor surveillance system, 2009–2013. *Prev Chronic Dis* 2016;13.
10. Althoff T, Sosič R, Hicks JL, *et al*. Large-Scale physical activity data reveal worldwide activity inequality. *Nature* 2017;547:336–9.
11. Thijs I, Fresiello L, Oosterlinck W, *et al*. Assessment of physical activity by wearable technology during rehabilitation after cardiac surgery: explorative prospective monocentric observational cohort study. *JMIR Mhealth Uhealth* 2019;7:e9865.
12. Cadmus-Bertram LA, Marcus BH, Patterson RE, *et al*. Randomized trial of a Fitbit-Based physical activity intervention for women. *Am J Prev Med* 2015;49:414–8.
13. Sun Y, Mobasheri A. Utilizing Crowdsourced data for studies of cycling and air pollution exposure: a case study using Strava data. *Int J Environ Res Public Health* 2017;14:274.
14. Sun Y, Du Y, Wang Y, *et al*. Examining associations of environmental characteristics with recreational cycling behaviour by Street-Level Strava data. *Int J Environ Res Public Health* 2017;14:644.
15. Stragier J, Mechant P, De Marez L, *et al*. Computer-Mediated social support for physical activity: a content analysis. *Health Educ Behav* 2018;45:124–31.
16. Chou W-YS, Prestin A, Kunath S. Obesity in social media: a mixed methods analysis. *Transl Behav Med* 2014;4:314–23.
17. 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;43:1575–81.
18. Zhang N, Campo S, Janz KF, *et al*. Electronic word of mouth on Twitter about physical activity in the United States: exploratory Infodemiology study. *J Med Internet Res* 2013;15:e261.
19. 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:e158.
20. Nguyen QC, McCullough M, Meng H-W, *et al*. Geotagged us Tweets as predictors of County-Level health outcomes, 2015-2016. *Am J Public Health* 2017;107:1776–82.
21. Joulin A, Grave E, Bojanowski P, *et al*. Bag of Tricks for Efficient Text Classification. arXiv:160701759 [cs] [Internet], 2016. Available: http://arxiv.org/abs/1607.01759 [Accessed 4 Feb 2019].

22. Harvard Health. Calories burned in 30 minutes for people of three different weights [Internet]. Harvard Health, 2017. Available: https://www.health.harvard.edu/diet-and-weight-loss/calories-burned-in-30-minutes-of-leisure-and-routine-activities [Accessed 26 Jun 2018].
23. Markham A. Fabrication as ethical practice. *Inform Commun Soc* 2012;15:334–53.
24. Longley PA, Adnan M, Lansley G. The Geotemporal demographics of Twitter usage. *Environ Plan A* 2015;47:465–84.
25. Burger JD, Henderson J, Kim G. Discriminating gender on Twitter. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsberg, PA, USA, 2011:1301–9.
26. Maurer D, Pathman T, Mondloch CJ. The shape of boubas: sound-shape correspondences in toddlers and adults. *Dev Sci* 2006;9:316–22.
27. Nielsen A, Rendall D. The sound of round: evaluating the sound-symbolic role of consonants in the classic Takete-Maluma phenomenon. *Can J Exp Psychol* 2011;65:115–24.
28. Mueller J, Stumme G. Gender inference using statistical name characteristics in Twitter. *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics*, Albany, NY, 2016:1–8.
29. Wolpert DH. Stacked generalization. *Neural Netw* 1992;5:241–59.
30. Nina C, Grant C, Nsoesie EO. "Detection of User Demographics on Social Media: A Review of Methods and Recommendations for Best Practices.". *ArXiv:1702.01807 [Cs]* 2017.
31. Cesare N, Grant C, Hawkins JB, *et al*. Demographics in social media data for public health research: does it matter? arXiv preprint arXiv:17101. *Bloomberg Data for Good Exchange Conference.*, New York, NY, 2017:1–8.
32. Centers for Disease Control and Prevention. Methods and references for County-Level estimates and ranks and State-Level modeled estimates, 2016: 2. Available: https://www.cdc.gov/diabetes/pdfs/data/calculating-methods-references-county-level-estimates-ranks.pdf
33. Faraway JJ. *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. Second Edition. CRC Press, 2016: 411.
34. Whitt-Glover MC, Taylor WC, Floyd MF, *et al*. Disparities in physical activity and sedentary behaviors among US children and adolescents: prevalence, correlates, and intervention implications. *J Public Health Policy* 2009;30(Suppl 1):S309–S334.
35. Wilson-Frederick SM, Thorpe RJ, Bell CN, *et al*. Examination of race disparities in physical inactivity among adults of similar social context. *Ethn Dis* 2014;24:363–9.
36. Valero-Elizondo J, Hong JC, Spatz ES, *et al*. Persistent socioeconomic disparities in cardiovascular risk factors and health in the United States: medical expenditure panel survey 2002-2013. *Atherosclerosis* 2018;269:301–5.
37. Dowda M, Ainsworth BE, Addy CL, *et al*. Correlates of physical activity among U.S. young adults, 18 to 30 years of age, from NHANES III. *Ann Behav Med* 2003;26:15–23.
38. Armstrong S, Wong CA, Perrin E, *et al*. Association of physical activity with income, Race/Ethnicity, and sex among adolescents and young adults in the United States: findings from the National health and nutrition examination survey, 2007-2016. *JAMA Pediatr* 2018;172:732-740.
39. Bancroft C, Joshi S, Rundle A, *et al*. Association of proximity and density of parks and objectively measured physical activity in the United States: a systematic review. *Soc Sci Med* 2015;138:22–30.
40. Trost SG, Owen N, Bauman AE, *et al*. Correlates of adults' participation in physical activity: review and update. *Med Sci Sports Exerc* 2002;34:1996–2001.
41. Centers for disease control and prevention (CDC). Neighborhood safety and the prevalence of physical inactivity-selected states, 1996. *MMWR Morb Mortal Wkly Rep* 1999;48:143–6.
42. Saelens BE, Sallis JF, Black JB, *et al*. Neighborhood-based differences in physical activity: an environment scale evaluation. *Am J Public Health* 2003;93:1552–8.
43. County HealthRankings [Internet]. Robert wood Johnson Foundation, 2016. Available: http://www.countyhealthrankings.org/homepage
44. US Census Bureau. Census regions and divisions of the United States, 2010. Available: https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
45. Martin SL, Kirkner GJ, Mayo K, *et al*. Urban, rural, and regional variations in physical activity. *J Rural Health* 2005;21:239–44.
46. Barker LE, Kirtland KA, Gregg EW, *et al*. Geographic distribution of diagnosed diabetes in the United States: a diabetes belt. *American Journal of Preventitive Medicine* 2011;40:434–9.
47. Barker LE, Kirtland KA, Gregg EW, *et al*. Geographic distribution of diagnosed diabetes in the U.S.: a diabetes belt. *Am J Prev Med* 2011;40:434–9.
48. Centers for Disease Control and Prevention. Early Release of Selected Estimates Based on Data From the 2016 National Health Interview Survey [Internet]. Available: https://www.cdc.gov/nchs/data/nhis/earlyrelease/earlyrelease201705.pdf [Accessed cited 4 Jun 2018].
49. World Health Organization. Physical Inactivity and Adults [Internet]. Global Strategy on Diet, Physical Activity and Health, 2019. Available: http://www.who.int/dietphysicalactivity/factsheet_adults/en/ [Accessed cited 11 Oct 2018].
50. Freeland AL, Banerjee SN, Dannenberg AL, *et al*. Walking associated with public transit: moving toward increased physical activity in the United States. *Am J Public Health* 2013;103:536–42.
51. Anderson M. Men catch up with women on overall social media use [Internet]. Pew Research Center, 2015. Available: http://www.pewresearch.org/fact-tank/2015/08/28/men-catch-up-with-women-on-overall-social-media-use/ [Accessed cited 3 Jul 2017].
52. Ainsworth BE, Caspersen CJ, Matthews CE, *et al*. Recommendations to improve the accuracy of estimates of physical activity derived from self report. *J Phys Act Health* 2012;9 Suppl 1:S76–S84.
53. Prince SA, Adamo KB, Hamel ME, *et al*. A comparison of direct versus self-report measures for assessing physical activity in adults: a systematic review. *Int J Behav Nutr Phys Act* 2008;5.
54. Sallis JF, Saelens BE. Assessment of physical activity by self-report: status, limitations, and future directions. *Res Q Exerc Sport* 2000;71(Suppl 2):1–14.
55. Van Dyck D, Cardon G, Deforche B, *et al*. Relationships between neighborhood walkability and adults' physical activity: how important is residential self-selection? *Health Place* 2011;17:1011–4.
56. Pinjari AR, Bhat CR, Hensher DA. Residential self-selection effects in an activity time-use behavior model. *Transport Res Part B Methodol* 2009;43:729–48.
57. McCormack GR, Shiell A. In search of causality: a systematic review of the relationship between the built environment and physical activity among adults. *Int J Behav Nutr Phys Act* 2011;8.
58. Boone-Heinonen J, Guilkey DK, Evenson KR, *et al*. Residential self-selection bias in the estimation of built environment effects on physical activity between adolescence and young adulthood. *Int J Behav Nutr Phys Act* 2010;7.
59. United States Census. Population Clock [Internet], 2019. Available: https://www.census.gov/popclock/ [Accessed cited 2019 Feb 21].
60. Pagoto S, Schneider KL, Evans M, *et al*. Tweeting it off: characteristics of adults who tweet about a weight loss attempt. *J Am Med Inform Assoc* 2014;21:1032–7.
61. Chunara R, Bouton L, Ayers JW, *et al*. Assessing the online social environment for surveillance of obesity prevalence. *PLoS One* 2013;8:e61373.
62. Salathé M, Khandelwal S. Assessing vaccination sentiments with online social media: implications for infectious disease dynamics and control. *PLoS Comput Biol* 2011;7:e1002199.
63. Chew C, Eysenbach G. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. *PLoS One* 2010;5:e14118.
64. Bernardo TM, Rajic A, Young I, *et al*. Scoping review on search queries and social media for disease surveillance: a chronology of innovation. *J Med Internet Res* 2013;15:e147.
65. Santillana M, Nguyen AT, Dredze M, *et al*. Combining search, social media, and traditional data sources to improve influenza surveillance. *PLoS Comput Biol* 2015;11:e1004513.