IMAGINE THE UNIVERSITY OF UTAH®

Twitter indicators and area-level health outcomes

Quynh Nguyen; Hsien-Wen Meng; Matt McCullough; Debjyoti Paul (University of Utah, Salt Lake City, UT)

Purpose

Build a national database with area level indicators of happiness and health behaviors from geotagged tweets. Compare Twitter-derived indicators with health outcomes at county and state levels.

Methods

79,848,992 million tweets were collected to build a new national data resource, HashtagHealth. We constructed indicators to capture mentions of popular food types and physical activities. A happiness score was assigned to each geo-tagged tweet. A total of 505,554 unique food-related businesses were collected from Yelp and spatially mapped. We test associations between social media variables and health outcomes at county and state levels.

Results

Table 1. Sentiment predictors of health outcomes, county level

	County-level predictors		
	Percent happy	Sentiment around healthy foods	Sentiment around physical activity
County-level health outcomes[a]	Beta (95% CI)[b]	Beta (95% CI)[b]	Beta (95% CI)[b]
All-cause mortality (per 100,000)	-7.37 (-13.89, -0.85)*	-3.39 (-6.97, 0.19)	-5.38 (-9.81, -0.94)*
Premature mortality (per 100,000)	-102.04 (-245.98, 41.90)	-87.79 (-178.45, 2.86)	-106.83 (-199.81, -13.85)*
Percent obesity	-0.67 (-1.11, -0.24)*	-0.42 (-0.60, -0.24)*	-0.53 (-0.82, -0.23)*
Percent diabetes	-0.10 (-0.25, 0.05)	-0.09 (-0.16, -0.01)*	-0.11 (-0.25, 0.02)
Percent physical inactivity	-0.75 (-1.18, -0.31)*	-0.55 (-0.76, -0.35)*	-0.62 (-0.91, -0.32)*
Percent poor/fair self-rated health	0.07 (-0.17, 0.30)	-0.10 (-0.23, 0.03)	-0.04 (-0.22, 0.15)
N	3117	2899	3054

[a]Data sources for health outcomes: 2011-2013 CDC WONDER mortality data; 2011-2014 Behavioral Risk Factor Surveillance System on adults aged 20 years and older

[b]Twitter variables were standardized to have a mean of 0 and standard deviation of 1. Adjusted linear regression models were run for each outcome separately. Models controlled for county-level demographics: median age, % non-Hispanic white, median household income. Standard errors accounted for clustering of county values at the state level

*p<0.05

Figure 1. National distribution of physical activity tweets, county level

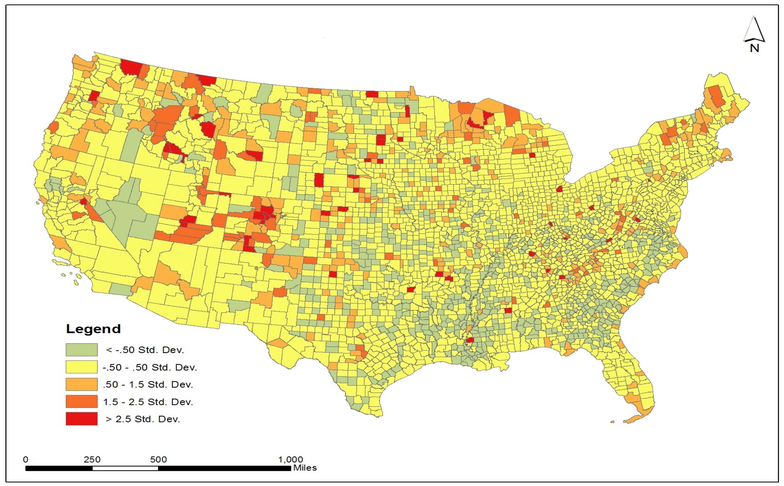


Table 2. State-level food environment and health outcomes

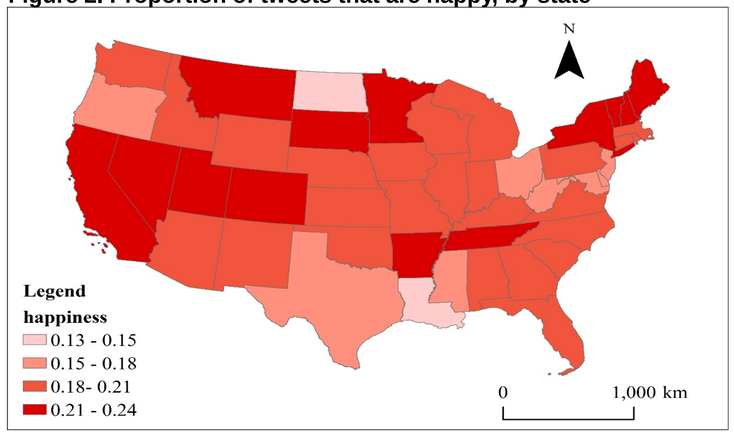
	State-level predictor variables[a]		
	Caloric density of Twitter food mentions	Percent Yelp listing, Café and bakeries	Percent Yelp listing, Burgers
State-level adult health outcomes	Beta (95% CI)[b]	Beta (95% CI)[b]	Beta (95% CI)[b]
All-cause mortality per 100,000	46.50 (25.81, 67.20)**	-31.06 (-48.69, -13.44)**	16.85 (-9.89, 43.59)
Percent diabetes	0.75 (0.42, 1.09)**	-0.66% (-0.92, -0.41)**	0.55% (0.14, 0.96)**
Percent prediabetes	-0.07 (-0.43, 0.28)**	0.05% (-0.22, 0.32)	-0.08% (-0.43, 0.26)
Percent obesity	1.78 (0.89, 2.67)**	-1.92% (-2.51, -1.32)**	1.35% (0.29, 2.40)*
Percent high cholesterol	1.40 (0.79, 2.00)**	-1.09% (-1.58, -0.60)**	0.36% (-0.43, 1.16)
Percent poor/fair self-rated health	2.01 (1.40, 2.61)**	-1.06% (-1.66, -0.45)**	1.12% (0.25, 1.98)*

[a]Predictor variables standardized to have a mean of 0 and standard deviation of 1. N=49. States in the contiguous United States, including District of Columbia

[b]Adjusted linear regression models were run for each outcome separately. Models controlled for state-level demographics: median age, % non-Hispanic white, median household income. Data sources for health outcomes: 2013 National Vital Statistics Reports, 2014 Behavioral Risk Factor Surveillance System

*p<0.05; **p<0.01

Figure 2. Proportion of tweets that are happy, by state



Findings in Summary

County-level

- Montana, Arizona, Wyoming, Utah, and Maine had the highest prevalence of physical activity mentions (Figure 1).
- Greater happiness levels was associated lower all-cause mortality, percent obesity, and percent physically inactive at county level (Table 1).
- Across 3000 US counties, Twitter indicators of happiness, food, and physical activity were associated with lower premature mortality, obesity and diabetes at the county level.

State-level

- A one standard deviation increase in caloric density of food tweets was related to higher all-cause mortality (+46.50 per 100,000) and higher prevalence of diabetes (+0.75%), obesity (+1.78%), high cholesterol (+1.40%), and fair/poor self-rated health (+2.01%) —controlling for state-level differences in age, percent non-Hispanic white, and median household income (Table 2).
- Higher percentage of tweets about alcohol and higher percentages of popular Yelp entries that were bars and pubs were related to higher state-level binge drinking and heavy drinking, but lower mortality and lower percent reporting fair/poor self-rated health

Conclusion

Social media represents a cost-efficient data resource for the capture of area-level socio-cultural characteristics. Twitter-derived data can be predictive of area-level health outcomes.

ACKNOWLEDGEMENTS: This work was supported by NIH grant 1K01ES025433-01 (Dr. Nguyen, PI)