Geotagged Tweets as Predictors of County-Level Health Outcomes

Quynh Nguyen, University of Utah; *Matt McCullough, University of Utah Department of Geography; Hsien-Wen Meng, University of Utah Department of Health, Kinesiology, and Recreation; Debjyoti Paul, University of Utah School of Computing; Dapeng Li, Michigan State University Center for Systems Integration and Sustainability; Suraj Kath, University of Utah School of Computing; Elaine Nsoesie, University of Washington Department of Global Health; James VanDerslice, University of Utah Department of Family and Preventive Medicine; Ken Smith, University of Utah Department of Family and Consumer Studies, and Huntsman Cancer Institute; Ming Wen, University of Utah Department of Sociology; Feifei Li, University of Utah School of Computing*

Background

Contextual factors can influence health through exposures to health-promoting and risk-inducing exposures. Nonetheless, the scarcity of consistently constructed contextual data limits understanding of contextual effects and geographical comparisons. Also, the environment is more than its physical features; social processes can affect health through the maintenance of norms, stimulation of new interests, and dispersal of knowledge.

Objective

Our aim was to build a national database from geotagged Twitter data with small-area indicators of prevalent sentiment and social modeling of health behaviors. We then examined whether Twitter characteristics predicted health outcomes.

Method

Between April 2015 and March 2016, we collected and spatially mapped 80 million publicly available geotagged tweets. We classified tweet sentiment using a Maximum Entropy classifier. Using a list of 1430 popular foods and 376 popular physical activities, we tracked the frequency of their social media mentions. In linear regression models, we used Twitter-derived indicators to predict health outcomes across 3000 US counties, controlling for county-level demographics and adjusting standard errors for clustering of county values at the state level. All variables were standardized to have a mean of 0 and standard deviation of 1.

Results

Higher percent happy (-0.07 SD), food (-0.14 SD), and physical activity (-0.12SD) tweets were associated with lower premature mortality. Higher prevalence of food tweets (B= -0.18 SD) and healthy food tweets (B= -0.09 SD) were associated with lower county-level obesity. Conversely, higher caloric density of Twitter food mentions (B= +0.08 SD) was related to higher county-level obesity. Higher prevalence of food tweets (B= -0.13 SD) and physical activity tweets (B= -0.12 SD) were related to lower county-level diabetes.

Conclusion

Social media represents a cost-efficient data resource for the construction of neighborhood features that, in turn, may influence community-level health outcomes.