

Project title: Representation and Bias in Self-Reported Physical Activity Across Social Media Platforms

Authors: Nina Cesare, Christan Grant, Elaine Nsoesie

Background:

There is broad and growing interest in the use of digital data as a source of data for public health research. Studies have used data from sources such as Twitter to assess sentiments toward vaccines (Bahk et al. 2016; Kang et al. 2017; Powell et al. 2016; Salathé and Khandelwal 2011), detect outbreaks of foodborne illness (Harris et al. 2014; Hawkins et al. 2016; Henly et al. 2017; Nsoesie, Kluberg, and Brownstein 2014), better understand postpartum health (De Choudhury, Counts, and Horvitz 2013) and sleep disorders (McIver et al. 2015), and track infectious disease such as influenza and dengue (Aramaki, Maskawa, and Morita 2011; Gomide et al. 2011; Paul, Dredze, and Broniatowski 2014). Distinct from studies that use online spaces as contexts in which to conduct studies of health behaviors, this literature leverages the unsolicited traces that individuals leave within digital contexts to better understand population-level health trends.

Regarding population health, digital data offer unique benefits for studying collective physical activity habits. For one, these data are geographically and temporally granular. They allow researchers to measure time as “continuous rather than bundled” (Savage, Ruppert, and Law 2010) and can be linked to small areas such as counties and tracts using geotags and self-reported locations. Additionally, research suggests that peer influence and self-comparison facilitated by social media platforms may impact individuals’ physical activity behaviors in positive ways, prompting individuals to use digital platforms to report activity and connect with others (Aral and Nicolaides 2017; Zhang et al. 2015, 2016).

This study seeks to assess representation and bias across social media platforms, using as a case study reports of physical activity. While the term ‘social media’ communicates the uniform experience of using digital platforms to connect with others, the way in which these platforms are used varies between and within each. This variation may stem from differences in platform affordances – defined as “the mutuality of actor intentions and technology capabilities that provide the potential for a particular action” (Faraj and Azad 2012; Majchrzak et al. 2013) – as well as users’ shared understanding of the platform’s functions and norms. Within sites, behavior may vary between users who occupy different demographic groups. It is known that this variation may introduce data bias that negatively impacts each platform’s ability to track public health trends, but no work has specifically assessed bias across platforms. To fill this gap, this project will compare the demographic composition of fitness-related digital trace data across select platforms and assess which source(s) of data most closely resemble the general population and most closely track population health trends.

Plan for Analysis:

Using social media sites Twitter, Instagram and social fitness tracker Strava as platforms, this project will accomplish two aims: (1) quantitative assessment of gender representation and (2) quantify spatial physical activity prevalence.

Aim 1: Quantitative assessment of gender representation. Under this aim, we will identify the types of activities reported within each platform, estimate the gender composition of individuals who self-report physical activity within each platform; and compare these measures to U.S. census demographic estimates at the county level. To assess gender on Instagram and Twitter, we will apply a previously

developed name-based classifier that leverages a weighted ensemble classification framework to predict whether a user is male or female with 82% accuracy. Gender information is self-reported on Strava. Finally, for data from Twitter and Instagram, we will use natural language processing to identify whether posts convey self-reported physical activity or mention activities in other contexts. Given past success using a feed forward neural network (FFNN) algorithm to predict relevant versus irrelevant physical activity tweets, we will begin by testing and refining this framework within our new data.

To assess representation, we will link geo-location data to U.S. counties and assess whether the gender distribution of the users from which we have data is significantly different than the distribution of the county from which the data are gathered. It is possible that women are more likely to discuss health via social media and therefore more likely to report physical activity. Assessing these patterns will help researchers account for data bias, as well as better understand how individuals use digital spaces for health and wellness.

Aim 2: Quantify Spatial Physical Activity Prevalence. Building on the results from Aim 1, we will use spatial regression models to ascertain whether physical activity prevalence from these digital platforms can be used to estimate actual prevalence, based on data from the U.S. Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor and Surveillance System Survey (BRFSS).

To accomplish this aim we will develop a series of mixed-effects regression models -a technique for modeling nested and/or geospatial data - to estimate US county-level physical inactivity based on measures of physical activity captured within each platform. Models will also control for basic sociodemographic features of each county (e.g. percent non-Hispanic white, median income) and built environment features that may impact collective health (e.g. percent with access to exercise spaces). These socioeconomic and environmental measures will be obtained from the United States Census' American Community Survey and the Robert Wood Johnson Foundation County Health Rankings dataset. Using these models, we will generate estimates of percent inactive within each county and assess how closely these estimates correlate with measures of inactivity from the CDC.

Discussion and Data Ethics

The goal of this project is to assess the representation and comprehensiveness of exercise as reported through select social media platforms. Findings will have implications for researchers interested in understanding bias in digital data used in public health research and applications. Furthermore, differences in representation and data content across platforms will be informative for social science researchers interested in understanding how individuals leverage digital tools for health and wellbeing.

For this project, we will gather and store data from Twitter, Instagram and Strava for tagged activities using each platform's public API. Following ethical guidelines of data use, we will not share user profile information with individuals outside of the research team and all raw data will be stored in a secure location. All data will be aggregated; individual data will be presented or published. A summary of this study will be submitted to the Institutional Review Board (IRB) at Boston University prior to analysis.

References:

- Aral, Sinan and Christos Nicolaides. 2017. "Exercise Contagion in a Global Social Network." *Nature Communications* 8:14753.
- Aramaki, Eiji, Sachiko Maskawa, and Mizuki Morita. 2011. "Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter." Pp. 1568–1576 in *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*. Stroudsburg, PA, USA: Association for Computational Linguistics.
- Bahk, Chi Y. et al. 2016. "Publicly Available Online Tool Facilitates Real-Time Monitoring Of Vaccine Conversations And Sentiments." *Health Affairs* 35(2):341–47.
- Cesare, Nina, Christan Grant, Jared B. Hawkins, John S. Brownstein, and Elaine O. Nsoesie. 2017. "Demographics in Social Media Data for Public Health Research: Does It Matter?" *ArXiv Preprint ArXiv:1710.11048*.
- Cesare, Nina, Quynh C. Nguyen, Christan Grant, and Elaine O. Nsoesie. 2018. "Estimating Obesity Prevalence by Age and Gender Using Social Media Data." Presented at the Population Association of America, April, Denver, CO.
- De Choudhury, Munmun, Scott Counts, and Eric Horvitz. 2013. "Predicting Postpartum Changes in Emotion and Behavior via Social Media." Pp. 3267–3276 in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Paris, France: ACM.
- Faraj, Samer and Bijan Azad. 2012. "The Materiality of Technology: An Affordance Perspective." *Materiality and Organizing: Social Interaction in a Technological World* 237–258.
- Gomide, Janaína et al. 2011. "Dengue Surveillance Based on a Computational Model of Spatio-Temporal Locality of Twitter." Pp. 3:1–3:8 in *Proceedings of the 3rd International Web Science Conference, WebSci '11*. New York, NY, USA: ACM.
- Harris, Jenine K. et al. 2014. *Health Department Use of Social Media to Identify Foodborne Illness — Chicago, Illinois, 2013–2014*. Center for Disease Control.
- Hawkins, Jared B. et al. 2016. "A Digital Platform for Local Foodborne Illness and Outbreak Surveillance." *Online Journal of Public Health Informatics* 8(1).
- Henly, Samuel et al. 2017. "Disparities in Digital Reporting of Illness: A Demographic and Socioeconomic Assessment." *Preventive Medicine* 101:18–22.
- Kang, Gloria J. et al. 2017. "Semantic Network Analysis of Vaccine Sentiment in Online Social Media." *Vaccine* 35(29):3621–38.
- Majchrzak, Ann, Samer Faraj, Gerald C. Kane, and Bijan Azad. 2013. "The Contradictory Influence of Social Media Affordances on Online Communal Knowledge Sharing." *Journal of Computer-Mediated Communication* 19(1):38–55.

- McIver, David J. et al. 2015. "Characterizing Sleep Issues Using Twitter." *Journal of Medical Internet Research* 17(6):e140.
- Nsoesie, Elaine O., Sheryl A. Kluberg, and John S. Brownstein. 2014. "Online Reports of Foodborne Illness Capture Foods Implicated in Official Foodborne Outbreak Reports." *Preventive Medicine* 67:264–69.
- Paul, Michael J., Mark Dredze, and David Broniatowski. 2014. "Twitter Improves Influenza Forecasting." *PLOS Currents Outbreaks*. doi: 10.1371/currents.outbreaks.90b9ed0f59bae4ccaa683a39865d9117.
- Powell, Guido Antonio et al. 2016. "Media Content about Vaccines in the United States and Canada, 2012–2014: An Analysis Using Data from the Vaccine Sentimeter." *Vaccine* 34(50):6229–35.
- Salathé, Marcel and Shashank Khandelwal. 2011. "Assessing Vaccination Sentiments with Online Social Media: Implications for Infectious Disease Dynamics and Control." *PLOS Computational Biology* 7(10):e1002199.
- Savage, Mike, Evelyn Ruppert, and John Law. 2010. "Digital Devices: Nine Theses." University of Manchester CRESC Working Paper 86. Retrieved from: <http://hummedia.manchester.ac.uk/institutes/cresc/workingpapers/wp86.pdf>.
- Zhang, Jingwen et al. 2016. "Support or Competition? How Online Social Networks Increase Physical Activity: A Randomized Controlled Trial." *Preventive Medicine Reports* 4:453–58.
- Zhang, Jingwen, Devon Brackbill, Sijia Yang, and Damon Centola. 2015. "Efficacy and Causal Mechanism of an Online Social Media Intervention to Increase Physical Activity: Results of a Randomized Controlled Trial." *Preventive Medicine Reports* 2:651–57.