EMORY UNIVERSITY SCHOOL OF MEDICINE

Injury Prevention Research Center at Emory

Publication Brief |June 2023

Can accurate demographic information about people who use prescription medications nonmedically be derived from Twitter?

One way of tackling the drug overdose epidemic is by improving surveillance methods. Traditional surveillance approaches, which include the National Survey of Drug Use and Health (NSDUH), overdose-related emergency department visits (EDV), and others, have considerable lags associated with the cycle of data collection, organization, and release.

Social media based surveillance has the potential of providing insights closer to real time. One major limitation of such methods is that it is not capable of providing insights by specific demographics or subpopulations. The objective of this study was to develop natural language processing methods to estimate the demographic distribution of people who mention nonmedical prescription medication use (NPMU) on Twitter.

Citation:

Yang, Y. C., Al-Garadi, M. A., Love, J. S., **Cooper, H. L. F.**, Perrone, J., & **Sarker, A.** (2023). Can accurate demographic information about people who use prescription medications nonmedically be derived from Twitter?. Proceedings of the National Academy of Sciences of the United States of America, 120(8), e2207391120. https://doi.org/10.1073/pnas.2207391120

Correspondence:

Abeed Sarker
abeed@dbmi.emory.edu

Data Collection Methods:

- Researchers developed a pipeline to collect publicly available twitter data to detect tweets mentioning prescription medications and self-reported NPMU from posts between March 6, 2018 to April 30, 2021.
- The gender, age, and race distribution estimation were performed by developing and validating natural language processing and supervised machine learning methods.
- The distributions were then compared with NSDUH 2019 and Nationwide Emergency Department Sample (NEDS) to validate the method.

Findings:

Automatically-derived statistics from Twitter closely resembled the metrics obtained from US Census data, particularly for race and gender identity. It has a lower proportion of females (4%) and Whites (1.5%), and more Hispanics (1.5%). In terms of age, Twitter has an over-representation of younger people compared with census estimates. Specifically, the proportion of people in the 18 to 25 group is approximately 10% higher, and the proportion for the 55+ group is 20% lower on Twitter compared with the census estimates.

Social media based surveillance for opioids, stimulants, and tranquilizers was consistent with statistics reported through traditional sources.

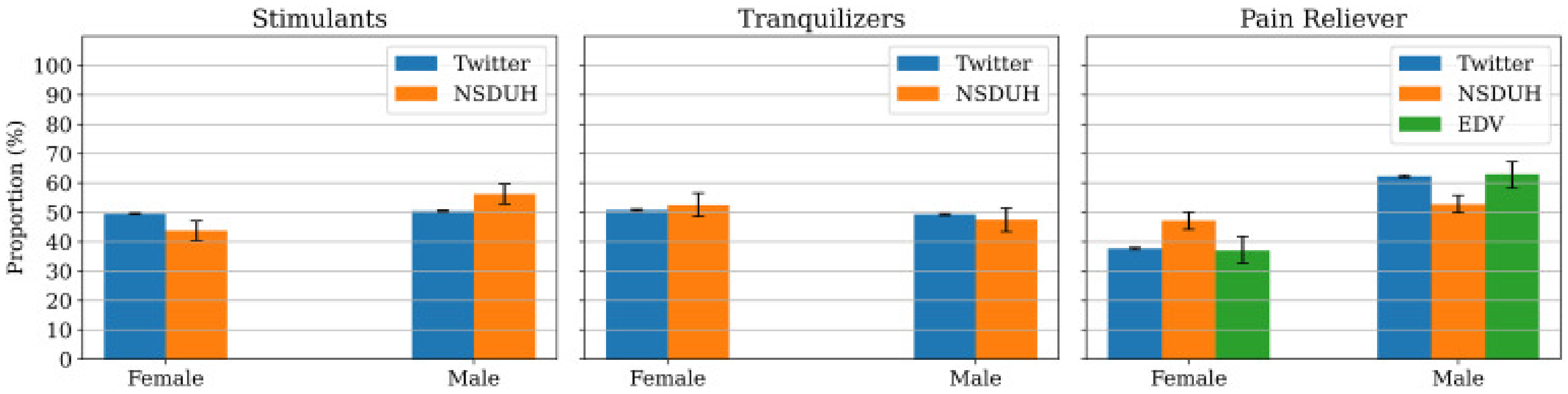

Figure 1. Gender distributions for NPMU estimated from Twitter and those reported in the NSDUH.

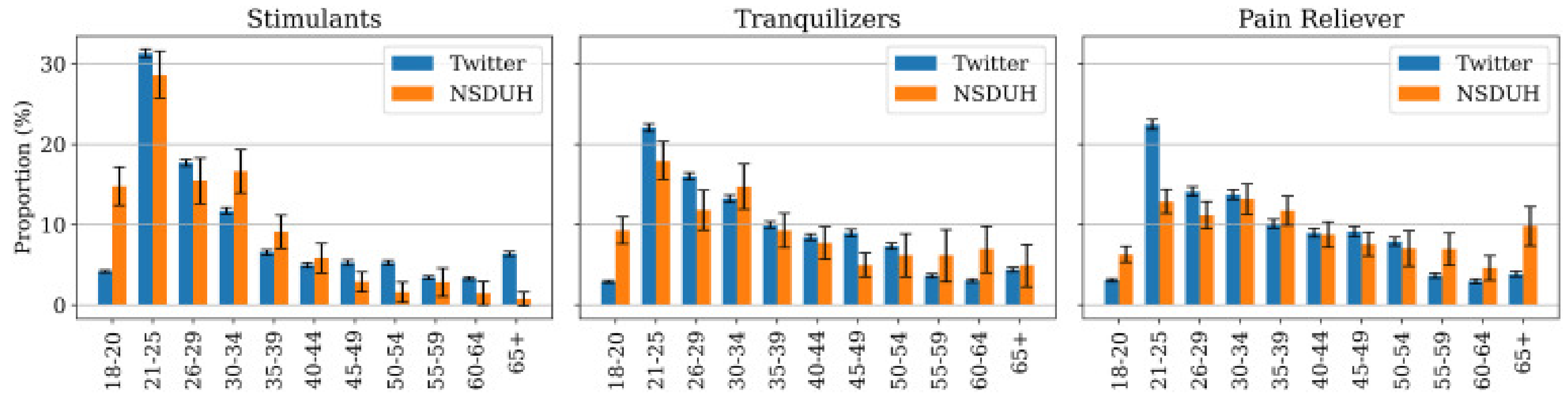

Figure 2. Age-group distributions for NPMU estimated from Twitter and those reported in the NSDUH.

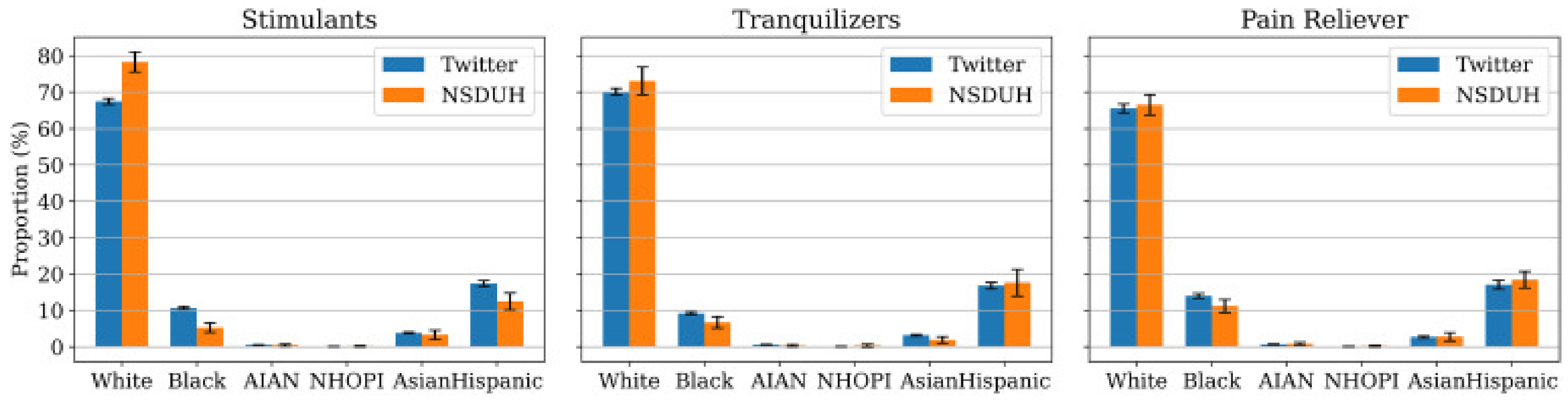

Figure 3. Race distributions for NPMU estimated from Twitter and those reported in the NSDUH.

Discussion:

A key advantage of the social media based automatic pipeline is that it can provide insights in close to real-time. Thus, it can complement traditional surveillance strategies and may help detect trends in substance use earlier. Limitations include errors made by the multiple natural language processing systems, and the skewed subscriber base of Twitter (over-representation of younger people). The researchers caution that the demographic estimations are only applicable on large samples and must not be used at the individual level. Researchers plan to improve the accuracies of their demography characterizations and make their methods more inclusive.