



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

Yuan-Chi Yang^a, Mohammed Ali Al-Garadi^{ab} , Jennifer S. Love^c , Hannah L. F. Cooper^d, Jeanmarie Perrone^e, and Abeed Sarker^{a,f,1}

Edited by Raminta Daniulaityte, The College of Health Solutions, Arizona State University, Phoenix, AZ; received May 4, 2022; accepted January 4, 2023 by Editorial Board Member Mary C. Waters

Traditional substance use (SU) surveillance methods, such as surveys, incur substantial lags. Due to the continuously evolving trends in SU, insights obtained via such methods are often outdated. Social media-based sources have been proposed for obtaining timely insights, but methods leveraging such data cannot typically provide fine-grained statistics about subpopulations, unlike traditional approaches. We address this gap by developing methods for automatically characterizing a large Twitter nonmedical prescription medication use (NPMU) cohort ($n = 288,562$) in terms of age-group, race, and gender. Our natural language processing and machine learning methods for automated cohort characterization achieved 0.88 precision (95% CI: 0.84 to 0.92) for age-group, 0.90 (95% CI: 0.85 to 0.95) for race, and 94% accuracy (95% CI: 92 to 97) for gender, when evaluated against manually annotated gold-standard data. We compared automatically derived statistics for NPMU of tranquilizers, stimulants, and opioids from Twitter with statistics reported in the National Survey on Drug Use and Health (NSDUH) and the National Emergency Department Sample (NEDS). Distributions automatically estimated from Twitter were mostly consistent with the NSDUH [Spearman r : race: 0.98 ($P < 0.005$); age-group: 0.67 ($P < 0.005$); gender: 0.66 ($P = 0.27$)] and NEDS, with 34/65 (52.3%) of the Twitter-based estimates lying within 95% CIs of estimates from the traditional sources. Explainable differences (e.g., overrepresentation of younger people) were found for age-group-related statistics. Our study demonstrates that accurate subpopulation-specific estimates about SU, particularly NPMU, may be automatically derived from Twitter to obtain earlier insights about targeted subpopulations compared to traditional surveillance approaches.

natural language processing | machine learning | Twitter | substance use | toxicovigilance

Substance use (SU), including nonmedical prescription medication use (NPMU), has been a major public health problem in the United States (US) for decades. Overdose deaths due to SU have steadily increased over the years, regardless of prevention measures (1). In 2020, the SU-related overdose death rate increased by 31% from 2019 to 28.3 per 100,000 population (2), over 20 times higher than the recorded rate in 1980 (1). In the 12 mo preceding March 2022, over 100,000 SU-related deaths are expected, as per provisional estimates, among the highest ever recorded in a 12-mo period (3). Due to the enormity of the SU epidemic, the US government and the White House have announced the deployment of unprecedented resources (4).

There are also significant disparities related to SU disorder (SUD) and the associated health outcomes. Many recent studies have highlighted the disparities depending on socioeconomic status, race/ethnicity, gender identity/biological sex, community, criminal history, and healthcare coverage (5–12). For example, studies have shown that non-Hispanic Blacks and Hispanics are less likely to receive buprenorphine treatment compared with Whites, and women are less likely than men (11, 13–15). Moreover, non-Hispanic Blacks and American Indians and Alaska Natives (AIAN) experienced the highest increases in the drug overdose mortality rates in 2019 and 2020 (16), while non-Hispanic Blacks experienced a much higher increase in mortality rates due to the coingestion of stimulants and opioids compared with non-Hispanic Whites (17). It has also been reported that people with lower income, living in non-metro urbanized regions, or who are uninsured are more likely to suffer from SUD (18). Multiple disparities may coexist, and exacerbate the likelihood of SU/SUD. Consequently, non-Hispanic Blacks, Hispanic/Latino persons, AIAN, and Native Hawaiian and Other Pacific Islanders (NHOPI), who also have low insurance coverage rates, face substantial SUD-related disparities (19). Distinct demographic groups may also have their own unique cultural and historical contexts and norms, consequently increasing the challenges associated with targeted surveillance and response.

Significance

Nonmedical substance use surveillance via social media has the potential to provide low-cost and more timely insights than traditional approaches. However, current social media-based approaches lack the capability to provide fine-grained, subpopulation-level statistics. We attempted to fill the gap by developing natural language processing methods to estimate the demographic distribution (gender, age, and race) of a large cohort ($N = 288,562$) of people who reported nonmedical prescription medication use on Twitter. Automatically derived distributions for opioids, stimulants, and tranquilizers were largely consistent, often with *very strong* correlations, with statistics reported in traditional sources such as the National Survey on Drug Use and Health. Our work represents an important stride in establishing social media as a complementary resource for substance use surveillance.

Preprint Server: medRxiv under a CC-BY-NC 4.0 International license (<https://www.medrxiv.org/content/10.1101/2022.04.27.22274390v1>).

Author contributions: Y.-C.Y., H.L.F.C., J.P., and A.S. designed research; Y.-C.Y., M.A.A.-G., and A.S. performed research; Y.-C.Y., M.A.A.-G., J.S.L., and A.S. analyzed data; and Y.-C.Y., M.A.A.-G., J.S.L., H.L.F.C., J.P., and A.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. R.D. is a guest editor invited by the Editorial Board.

Copyright © 2023 the Author(s). Published by PNAS. This open access article is distributed under [Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 \(CC BY-NC-ND\)](https://creativecommons.org/licenses/by-nc-nd/4.0/).

¹To whom correspondence may be addressed. Email: abeed@dbmi.emory.edu.

This article contains supporting information online at <https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2207391120/-/DCSupplemental>.

Published February 14, 2023.

An important aspect of effectively tackling the drug overdose epidemic and alleviating disparities is to improve surveillance, specifically to accelerate the data curation process to provide timely, accurate, and actionable insights (20, 21). Traditional surveillance approaches and/or sources of data include surveys, such as those conducted by the National Survey on Drug Use and Health (NSDUH) (22), poison control centers (23), hospital data about treatment admissions and discharge (24), overdose-related emergency department visits (EDV) (25, 26), and overdose death records (27). Such traditional surveillance systems have considerable lags associated with the cycle of data collection, organization, and release. For example, the 2020 NSDUH Annual National Report was not available until the end of October 2021. Due to such lags, trends in SU/overdose are only detected and understood retrospectively, often after considerable damage has already been done and/or SU patterns have shifted. The lag is particularly problematic since the SU/overdose epidemic has been continuously evolving over the years. For example, the primary contributor to overdose-related deaths in the early 2000s (US) was cocaine, which was later taken over by prescription opioids followed by heroin (1). Also, in recent years, there have been notable increases in deaths due to synthetic opioids (e.g., fentanyl) and psychostimulants (e.g., methamphetamine) (27–30). Reliance on traditional surveillance approaches means that the exact current trajectory of the epidemic will only be known months from now. Therefore, there is an urgent need for accurate, close-to-real-time surveillance systems for SU.

To address the shortcomings of traditional approaches, social media have been proposed as potential complementary resources for timely surveillance (31–33). Over 220 million Americans (~70% of the population) use social media, and many discuss health-related topics. Discussions of health-related topics include self-reported SU and SUD. Theoretically, these publicly available discussions can be mined in close to real-time using natural language processing (NLP) methods. However, NLP of health-related chatter, including SU-related chatter, is hard due to various characteristics of the data, such as the presence of colloquial expressions, misspellings, and noise. In the space of SU/SUD research from social media, researchers have developed and applied progressively sophisticated methods over the last decade. Early research attempted to leverage data from online health communities that had dedicated forums for discussions about nonmedical use (NMU). For example, MacLean et al. studied data from an online health community named Forum77 to investigate the efficacy of online mutual help groups for NPMU associated with opioids (34). Relying on online health communities with dedicated forums ensures that studies have access to rich data, although the volume of information may be low since the subscriber bases of such communities are not very large. Studies utilizing data from generic social networks such as Twitter initially focused primarily on deriving insights based on the volume of chatter about specific substances. Hanson et al. (35), for example, collected posts mentioning “*adderall*” from Twitter, and demonstrated that the volume of posts substantially increases during months when college students have their examinations since many such students nonmedically use stimulants to enhance performance. Graves et al. (36) and Chary et al. (37) combined volume-related statistics with geolocation metadata on Twitter to demonstrate that the volumes of opioid-related chatter had some correlations with statistics derived from traditional sources such as overdose death rates. Sarker et al. (38), showed that only a minority of NPMU chatter on social media are first-person reports, and they proposed a supervised classification strategy

to filter out noise and build cohorts of people who report NPMU. Correlations with geolocation-specific metrics from traditional sources have been shown to be stronger once a supervised classification filter is applied (39). While a number of studies have been able to leverage metadata accompanying social media posts, such as those from Twitter, to obtain geolocation-specific insights, it has not been possible to group insights based on other demographic characteristics. This poses a barrier to conducting fine-grained, subpopulation-specific research using such data—a clear disadvantage compared with the NSDUH and other traditional sources. Ideally, SU surveillance data need to cover the full range of demographics (e.g., race, age, gender, and geographical area), and contain sufficient granularity to observe subtle differences among different demographic groups. Methods for accurately and automatically estimating the distributions of key demographic features in social media subscriber cohorts can enable fine-grained subpopulation-level analyses and comparisons—a gap we attempt to address in this paper.

We describe the development and validation of methods for automatically estimating demographic distributions (age-group, gender, and race) in a Twitter cohort consisting of subscribers who self-reported NPMU. We integrated the methods that we developed to establish an end-to-end data-centric cohort characterization pipeline and applied it to the Twitter NPMU cohort. To validate our pipeline, we compared the distributions estimated from Twitter with those reported in traditional sources [NSDUH 2019 (18) and Nationwide Emergency Department Sample (NEDS) (40)] for prescription stimulants, tranquilizers, and opioid pain relievers. Due to the absence of any prior work on this specific topic, our objectives were to discover the trends in a purely data-centric manner. That is, we attempted to understand the similarities and differences between our findings and the statistics reported in traditional sources, rather than testing the hypothesis that our system can reproduce the same results in the traditional sources or replace the traditional methods.

Results

Twitter NPMU Cohort. We collected tweets mentioning prescription medications and detected self-reported NPMU using a supervised classification system that we developed and optimized in our prior work (41). Posts were collected from March 6, 2018 to April 30, 2021. Our system detected 482,902 NPMU-indicating tweets that were posted publicly and extracted their authors’ metadata, including post history, if available. In this manner, we collected the metadata of 288,562 Twitter subscribers who posted the NPMU-indicating tweets, and their past posts (over 1 billion tweets). We refer to this cohort-level dataset as the Twitter NPMU cohort.

Gender, Age, and Race Distribution. The gender, age, and race proportions for Twitter subscribers, estimated from the 2018 Twitter Survey conducted by the Pew Research Center (42), and those reported in the US Census (18) are shown in Fig. 1. The estimated gender and race proportions from the two sources are comparable, while the age proportions are substantially different. Compared with the US Census data, Twitter has marginally lower proportions of females (4% less) and Whites (1.5% less), and more Hispanics (1.5% more) (43). The closeness of the proportions from the two sources suggests that Twitter-based estimates specific to gender and race may be representative of the country’s population. In contrast, in terms of age, Twitter has an

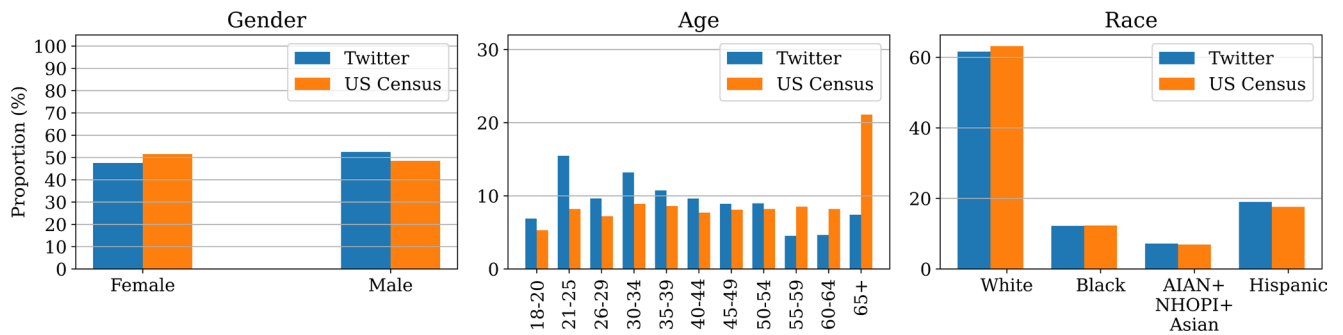


Fig. 1. Gender, age, and race proportions estimated from Twitter and those reported in US census.

overrepresentation of younger people compared with the 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. The overrepresentation of younger people on Twitter, and social media in general, is a well-known phenomenon. In terms of NPMU, the distributions automatically estimated from Twitter were mostly consistent with the statistics from NSDUH and NEDS, with 34 out of 65 (52.3%) estimates falling within the 95% CIs of the metrics reported in the latter sources. We provide further details below.

Gender Distribution Estimates. The estimated gender proportions from Twitter data and the gender proportions from traditional sources, including NSDUH and NEDS, are given in Fig. 2 (further details in *SI Appendix, Table S1*). The three categories of medications included were opioid pain relievers, tranquilizers, and stimulants. For NPMU of tranquilizers, the estimated Twitter proportions are within the 95% CIs of the proportions reported in the NSDUH. For stimulants, the Twitter proportion estimate for females is slightly higher than the NSDUH reported number (~5%). For opioids, the estimated proportions are significantly different between Twitter and the NSDUH (~10%). Specifically, the proportion of females on Twitter is lower than the NSDUH estimates. Interestingly, however, we found that the numbers reported by the NSDUH also differ in terms of proportions from the opioid-related EDVs reported in NEDS, but the estimates from the latter are very close to the Twitter proportions (no significant difference). This suggests that estimates derived from Twitter may be more reflective of overdose-related events rather than NMU for this category. Statistical significance in correlation between Twitter and NSDUH could not be established because of small N (Spearman r : 0.66; P = 0.27).

Race Distribution Estimates. The estimated race proportions from Twitter and the proportions from the NSDUH are shown in Fig. 3 (further details in *SI Appendix, Table S2*). The estimated distributions are similar between the two data sources, and each proportion from Twitter is either within or close to the 95% CI of the corresponding proportion reported in the NSDUH. For all medication categories, the majority of people who reported NMU are White, followed by Hispanic and Black. Asians who reported NMU only represent about 4% or less of the cohort; AIAN and NHOPI groups each represents less than 1% of the cohort. Importantly, the Twitter data had representation from all of the minority races. The most prominent differences between Twitter and the NSDUH are for the White and Hispanic stimulant groups, and for Blacks across all medications (Twitter estimates are higher than NSDUH). Overall, the Twitter estimates are very strongly correlated with the NSDUH statistics (Spearman r : 0.978; P < 0.005).

Age-Group Distribution Estimates. The estimated age-group proportions from Twitter data and the proportions from the NSDUH are shown in Fig. 4 (details in *SI Appendix, Table S3*). For most age groups, the Twitter and the NSDUH estimates are similar, and the overall correlation is strong (Spearman r : 0.673; P < 0.005). The most prominent differences are for young adults (18 to 20 and 21 to 25) and the elderly (65+). The estimated proportions from Twitter are consistently lower for the 18 to 20 group and higher for the 21 to 25 group. For the 65+ group, the estimated Twitter proportion is higher for stimulants, similar for tranquilizers, and lower for pain relievers compared with the NSDUH numbers. For opioid pain relievers, the estimated Twitter proportion for the 21 to 25 group is approximately 10% higher, and for the 65+ group, 6% lower compared with the NSDUH. For tranquilizers, the estimated Twitter proportion for the 18 to 20 group is approximately 6% lower compared with the NSDUH,

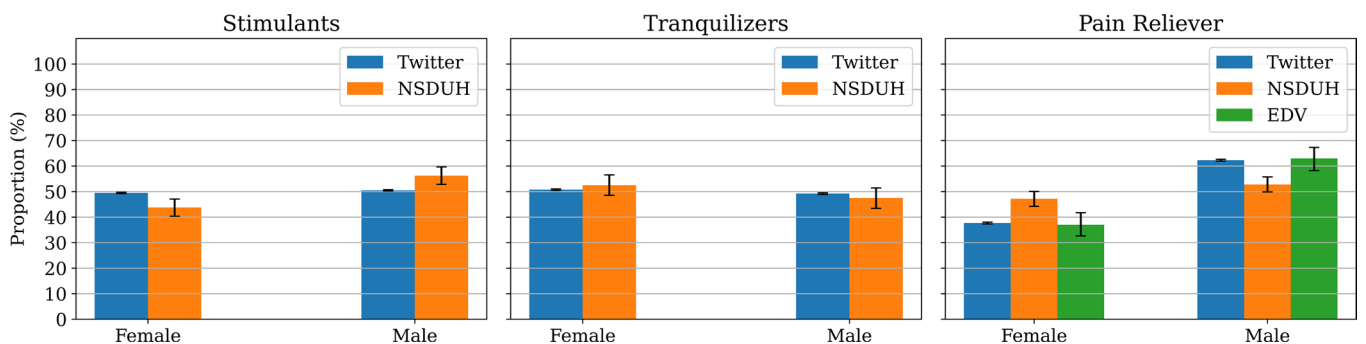


Fig. 2. Gender distributions for NPMU estimated from Twitter and those reported in the NSDUH. For opioid pain relievers, the gender distribution of overdose-related emergency medicine visits is also provided. 95% CIs are provided for each bar.

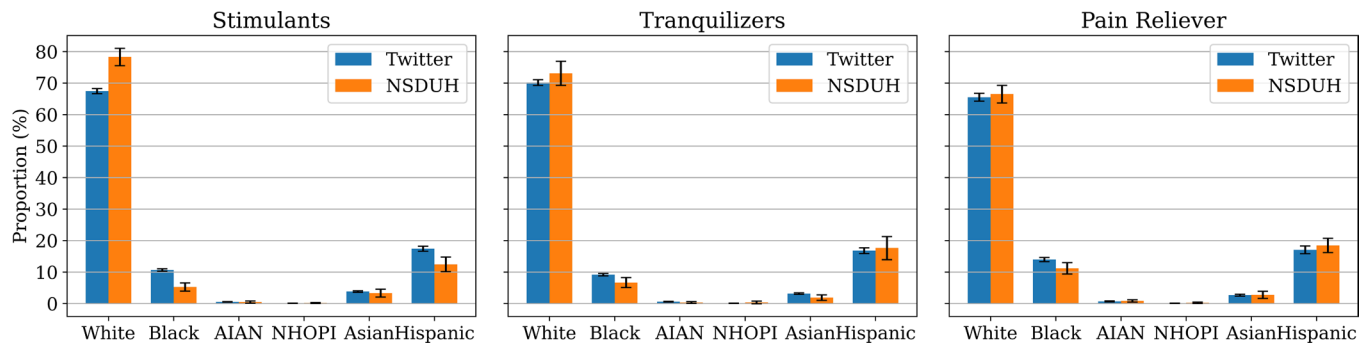


Fig. 3. Race distributions for NPMU estimated from Twitter and those reported in the NSDUH. 95% CIs are provided for each bar.

but the proportions for the 65+ group are not significantly different. For stimulants, the estimated Twitter proportion for the 18 to 20 group is approximately 10% lower, and for the 65+ group 6% higher compared with the NSDUH. We have also compared the age-group proportions from Twitter data for the NPMU of pain relievers with the proportions from the opioid-related EDVs reported in NEDS. The proportions are shown in *SI Appendix, Fig. S1 and Table S4*. We found that the Twitter and the EDV estimates are mostly similar (Spearman r : 0.943; P = 0.005), with prominent differences for young adults (20 to 24).

Discussion

Our current work automatically estimates the distribution of demographic characteristics in a large Twitter cohort—in this case, a cohort of subscribers who self-reported NPMU—and compares the automatically obtained distributions with those reported in traditional sources. Our experiments validate that most of the estimates derived from Twitter are consistent with those reported in traditional sources, such as the NSDUH and NEDS. The major differences were in the age-group-based estimates, specifically for young adults (18 to 20, and 21 to 25) and the elderly (65+).

The NSDUH is conducted as a survey among the noninstitutionalized population in the US and thus is limited by the respondents' truthfulness and exclusion of individuals in hospitals, prisons, or even treatment centers (44). It is reported that the respondents tend to underreport or overreport on surveys, and this tendency is influenced by their demographics including gender, race, or age (45–49). For example, among surveyed cocaine users, African American, young adults (18–30), and females were found to be more inclined to underreport (45). Though Twitter data also rely on individual subscribers' truthfulness and willingness to share, we suspect that the default anonymity of Twitter accounts partially mitigates demography-specific underreporting than the NSDUH

and, thus, might even be better suited for analyzing subpopulation differences than the NSDUH. We speculate that the closeness of the gender and race distributions for Twitter subscribers and the US population, as depicted in Fig. 1, is a key reason for the Twitter-based estimates to be very close to the NSDUH, while the differences might be explained by the under/overreporting tendencies. For example, the underreporting tendency of females for cocaine might help explain the apparent overestimation of female stimulant users on Twitter, and the underreporting tendency of African Americans might be crucial to understand the overestimation of African American users for all three medication categories on Twitter (45). Similarly, the different tendencies of under/overreporting and Twitter usage among distinct age groups might contribute to the differences in the age-wise estimates. Also, though Twitter data are limited to those who have internet access, it may capture a certain portion of the institutionalized or before/after their institutionalized periods. It is even possible that Twitter is less biased against the incarcerated Black population than the NSDUH.

Evaluation of Performance as a Surveillance System. Our pipeline is advantageous in its timeliness, flexibility, simplicity, and stability. The data collection can be done continuously and in near real-time on a personal desktop with internet connection while requiring minimal human supervision. Instead of structured answers to questionnaires, our collected data contain salient unstructured text information, allowing data mining with research questions evolving over time. For example, the collected tweets contain information regarding how and why the authors are using the substances. Collecting similar information from surveys usually requires incorporating prior knowledge into question design, but that is not necessary for our pipeline. However, since the data are usually massive, the data analysis is typically done using NLP and machine-learning methods. The advantage is that,

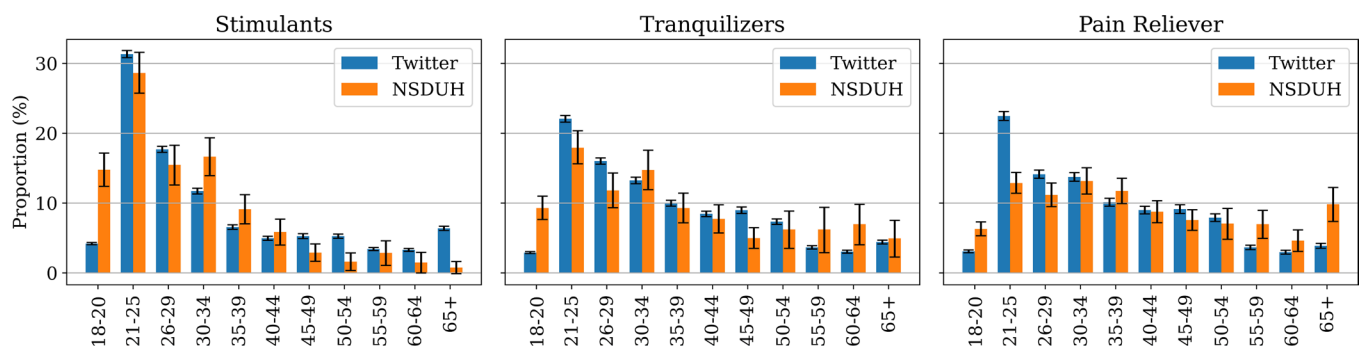


Fig. 4. Age-group distributions for NPMU estimated from Twitter and those reported in the NSDUH. 95% CIs are provided for each bar.

once the tailored scripts are developed, they can often be run automatically and on the fly. Expanding the pipeline (e.g., to collect cohorts of targeted illicit substances) would only require a university informatics team to dedicate a few months from initial exploration to operational prototypes. These advantages of our system fit perfectly into the CDC's Data Modernization Initiative, Priority 2—*Accelerate Data into Action to Improve Decision-Making and Protect Health*.

Data quality, acceptability, sensitivity, positive predictive value (precision), and representativeness of our pipeline are inevitably limited by Twitter's subscriber base and the subscribers' willingness to share their information publicly and truthfully. Demographic information is often not disclosed, and we have no method to validate the users' claims. Notably, our data do not rely on memory as much as surveys, as the time leap is limited to the time of NPMU to the time of posting. Also, we speculate that the Twitter subscribers might be more truthful in their self-disclosures because they have more power to choose what they are willing to share publicly. Therefore, though our method of social media listening does have disadvantages, we believe it is an important complementary surveillance system with high potential.

Related Work. Earlier, we provided a chronological review of related literature focusing on social media mining for SU surveillance, so in this subsection, we focus our review on demographic information detection and analyses from social media data. Our work is not the first that aims to detect demographic information from social media, although it develops and applies such methods for characterizing a specific large cohort. Past studies have proposed cohort characterization methods, including gender (50–53), age (52, 54–56), and race (57–60). Typically, these pipelines comprised supervised classification methods and used subscribers' metadata including names, usernames, bio, past tweets, or even images as features. Among these, the gender detection methods were reported to be the most accurate, with classifiers achieving accuracies above 94% (61). In contrast, race and age estimation pipelines proposed in past research had not obtained high accuracies. They also often do not provide sufficient granularity. The race/ethnicity estimation pipelines reported in the literature usually focus only on four categories (White, Black, Asian, and Hispanic/Latino) or less, leaving out AIAN and NHOPI (57–60), although AIAN has the highest overdose mortality rate among all race groups (16, 62). For age detection, the groupings often do not match those defined in the NSDUH, making comparative evaluations impossible (54–56). Though regrouping is possible for a few methods, they were not developed based on Twitter and thus may have limited applicability (52). In our work, we developed age and race estimation pipelines with high precision and fine granularity (11 for age and six for race) based on our Twitter data. Due to the paucity of annotated datasets, we applied a search-based approach that employs text pattern matching for detecting self-disclosed age and race. Because there is no gold standard for the negative case (i.e., subscribers who have not self-disclosed their age or race using the specified pattern), we focused on improving precision while maintaining an acceptable retrieval rate. Precision is preferred over recall (i.e., some cohort members will be missed) since the number of Twitter cohort members is large and growing over time, so obtaining sufficient numbers of people from each category is not a bottleneck. Specific details about our developed algorithms are provided in the *Materials and Methods* section.

Potential Applications and Future Work. There are several aspects of our automatic pipeline that can be improved in the future. First, we may be able to improve the performance of

the classifier that detects the self-disclosures of NPMU. One potential route is to examine if the classifier underperforms for any demographic group due to different self-disclosure behaviors, and then fine-tune the classifier accordingly. Second, we may make our findings more reliable by further improving the age group and race characterization methods. Three potential directions include i) annotating more tweets matched by the text patterns (i.e., creating a larger gold standard), ii) enriching the set of the text patterns, and/or iii) replacing the rule-based method with a machine learning-based classifier. Third, the pipeline can be extended to illicit substances, including opioids such as heroin, and stimulants such as methamphetamine, enabling its use as a more comprehensive surveillance system.

While our work is a significant stride toward moving social media from a fringe resource to an important one for SU surveillance, there are further opportunities for future work. As mentioned, extending our methods for including illicit substances can be an important step. The literature suggests that people who report NPMU may also be more likely to be exposed to illicit substances (63). The ability to detect self-reported illicit SU combined with our ability to collect and analyze longitudinal data about cohort members may enable us to detect common trends in the transition from prescription to illicit substances (e.g., from oxycodone to fentanyl) and vice versa. Longitudinal data collection and analysis from the automatically curated cohort may also enable us to detect novel psychoactive substances that infiltrate these communities (e.g., designer benzodiazepines) faster. Compared with the cross-sectional nature of NSDUH, our approach enables longitudinal analyses of targeted subpopulations that have often been underrepresented in traditional studies. Exacerbated by the war on drugs and inequalities in enforcement and incarceration, some of these groups are extremely hard to reach by traditional means, and consequently, there is now great interest in studying data from such groups (64–66). The ability to reach these hard-to-reach subpopulations and longitudinally track them may also open up opportunities for targeted interventions, an area that has not been explored in past research. Interventions, for example, could focus on actively discouraging high-risk cohort members from engaging in life-threatening behavior or connecting them to their local harm reduction services proactively, rather than waiting for them to initiate contact.

Limitations. Our work is largely limited by the data source. The demographic distribution of Twitter subscribers is different than the US population, especially for certain age groups. Though we adjust the estimates according to the 2018 Twitter survey, it only partially solves the issue. For people who do not use Twitter or do not discuss their NMU, we have no other means to collect their data and, thus, there may be groups of people that are not represented in our analyses. Also, social media data are noisy and may contain false information (e.g., fake races or genders), which we have no alternative approach to verify. The reliance on a large number of cohort members, however, somewhat mitigates this limitation as the impact of small amounts of fake information is likely to be minimized or removed. Additionally, tweets are short (limited by 280 characters) and consist of colloquial words, posing significant limitations on the classifier development, and thus the performance of the overall pipeline.

Our proposed methods *cannot* and *should not* be used for identifying the age group, race, or gender identity of individual Twitter subscribers. Their performances are not perfect; so there is no guarantee that they will not incorrectly characterize a single subscriber. Our methods essentially *estimate* the distributions of demographic information within a given *cohort*. Due to the large

size of the cohort, we anticipate that the small numbers of incorrect characterizations are eclipsed.

Finally, an important limitation of our current work is that it excludes certain segments of the population who have been underrepresented in past research—such as those with nonbinary gender identities. Our current work could not achieve such level of granularity due to lack of available data. The promising findings from our current work is the first step and our planned future work includes improving the inclusivity of our analysis (e.g., by inclusion of nonbinary population, the uninsured, and those unreachable via traditional means).

Materials and Methods

Twitter NPMU Cohort. Twitter data were collected through a data processing pipeline that we have described in past research (41, 67). The components of the pipeline include collecting publicly available streaming data about prescription medications, classifying the data, and then retrospectively collecting the metadata of subscribers who are detected to self-report NPMU. We collected English tweets mentioning at least one of over 20 PMs (including generic and trade names, and common misspellings) (68) that have the potential for NPMU (see *SI Appendix, Table S5* and "*Note on keywords used for data collection*"). The full set of keywords used for data collection includes about 600 terms and are given in Supplementary Files, "keywords-1.json" and "keywords-2.json" (67). We developed annotation guidelines with our domain expert (J.P.), and annotated a large subset of 16,443 tweets into four categories: NPMU (tweets indicating that the subscribers have nonmedically used the medications or intended to do so), consumption (tweets indicating that the subscribers have used the medications and containing no sign of NPMU), information (tweets that only mention the medications but do not indicate any use), or nonrelevant (usually tweets that mention the search keywords but refer to things other than the medications) (67). Please see *SI Appendix, Tables S6 and S7* for a more detailed description of the NPMU class and examples. Please also see our previous articles dedicated to the annotations, O'Connor et al. (2020), for more details (67). We used the annotated data to train machine-learning classifiers (including SVM, Random Forest, Bidirectional Long-Short Term Memory, and several transformer-based models), and the best performing classifier is based on RoBERTa-Large (69), a transformer-based model, with an accuracy of 82.3% (41). We then deployed the best-performing classifier to classify the streaming data into the four categories (including NPMU). For the subscribers whose tweets were classified as NPMU by the machine-learning classifier, we extracted all their available Twitter data, including subscriber profiles and past tweets (excluding retweets). We continued to collect their past tweets every 2 wk up to the date of data extraction for this study. These subscribers make up our NPMU cohort.

Gender Estimation. The genders of the Twitter subscribers were estimated based on the meta-1 classifier in our prior work, which is based on an SVM classifier trained on subscribers' past tweets and the M3 classifier trained on subscribers' names, screen names, and descriptions (bios) (53, 61). The genders are treated in a binary framework (excluding those with nonbinary gender identities due to lack of available data) and should be interpreted as how subscribers represent themselves online and thus are closer to the subscriber's gender identities than biological sexes. We developed the classifier based on gender-labeled datasets made available by Liu and Ruths (51) and Volkova et al. (70). In total, we were able to retrieve the metadata of 67,181 subscribers, consisting of 35,812 (53.3%) females and 31,369 (46.7%) males, which we used to develop the pipeline. We validated the performance on a set of 412 subscribers in our NPMU cohort whose genders were identified using the public gender fields on their linked Facebook account profiles. The classifier achieved accuracy of 94.4% (95% CI: 92.0 to 96.6) on this set.

Age Estimation. The ages of the Twitter subscribers were estimated based on a rule-based approach that optimizes precision (positive predictive value). Our pipeline searches for text patterns that are self-disclosures about the subscribers' ages. The pattern matching is done using regular expressions. Sample text patterns include "(d\d) birthday to me" or "i'm (d\d)" where "d" denotes digits (0 to 9). We also constructed a filter to remove irrelevant statements that are

not associated with ages, such as "I'm 20 weeks pregnant." For subscribers who have multiple tweets matched, we constructed a rule-based module to detect potential fraudulent information and infer the subscribers' ages. The pipeline was developed based on a set of 2,000 subscribers, among which 1,540 tweets from 609 subscribers matched the text patterns and were annotated. The annotation agreement based on an overlapping set of 346 subscribers (952 tweets) is 89.3% with Cohen's $K = 0.89$ (95.8% with Cohen's $K = 0.96$ on tweets). The test accuracy is 0.88 (95% CI: 0.84 to 0.92) [0.90 (95% CI: 0.86 to 0.94) when allowing a 1-y age discrepancy] on the subscribers who have matched tweets (referred to as precision in the text) and 0.93 (95% CI: 0.90 to 0.95) on the matched tweets. The age estimation pipeline can be found as `ageCharacterization.py` in Supplementary File "script.rar."

Race Estimation. The race estimation module is similar to the age estimation one and applies rules and patterns. We consider the following race classification: White, Black, Asian, Hispanic, American Indian and Alaska Native (AIAN), and NHOPI. Relevant expressions indicating race are searched using regular expressions. Example text patterns include "i'm (black)" or "i'm (white)". We also constructed a filter to remove irrelevant statements such as "I'm black salmon." For subscribers who have multiple tweets matched, we constructed a rule-based function module to detect potential fraudulent information and estimate the subscribers' races. The pipeline was developed based on a set of 4,000 tweets, among which 1,124 tweets from 578 subscribers matched the text patterns and were annotated. The annotation agreement based on an overlapping set of 293 subscribers (533 tweets) is 87.7% with Cohen's $K = 0.78$ (94.0% with Cohen's $K = 0.88$ on tweets). The test accuracy of the pipeline is 0.90 (95% CI: 0.85 to 0.95) on the subscribers who have matched tweets (referred to as precision in the text) [0.94 (95% CI: 0.91 to 0.97) on the matched tweets]. The race estimation pipeline can be found as "raceCharacterization.py" in Supplementary File "script.rar."

Gender, Age, and Race Distribution Estimation. The gender, age, and race distribution estimation are performed by applying the gender classifier, the age pipeline, or the race pipeline to the NPMU cohort. The gender distributions are estimated using 288,562 NPMU subscribers whose genders could be inferred. The age distributions are estimated using 63,073 NPMU subscribers whose age could be inferred. The race distributions are estimated using 32,784 NPMU subscribers whose race could be inferred. Since the race pipeline is not designed for subscribers who have more than one race ("more" in the 2018 Twitter Survey (42) and the NSDUH (18)), we did not include those subscribers when reporting the results and comparing them with the references.

Baseline for Twitter Subscribers' Demographics. We established the baseline for the US Twitter subscribers based on the 2018 Twitter Survey conducted by Pew Research Center (42). The raw data can be obtained as a .sav file by requesting the access from the Pew Research Center. We focused on Twitter subscribers who have at least used Twitter once a week (the value for field "TWITTER_USE" is four or smaller) and calculated the proportions of the targeted demographics (e.g., age, gender, and race) among these subscribers. We used the field "PPAGE" for age, "PPGENDER" for gender, and "PPETHM" for race. We noted that the survey was conducted only among the adults (aged 18+) and the Asian, AIAN, and NHOPI races were grouped into "others." Therefore, all the age groups including age below 18 are dropped.

Baseline for Twitter Age and Race Characterization. We established the baseline for Twitter Age and Race characterization based on a dataset of 156,368 general Twitter subscribers. We first collected streaming tweets with the English stopwords in the Natural Language Toolkit (NLTK) as keywords on Aug 27, 2021 using Twitter API (71). The NLTK is a widely used python package for NLP. Stopwords is a list of words that are commonly used but do not carry domain-specific meaning and are often dropped during text mining. Some examples in English including "I", "you", "is," ... etc. Since these words appear very frequently in text, including Twitter posts, such a data collection scheme can collect an approximately random set of Twitter subscribers. We then collected the subscribers' metadata and applied the age and race pipelines. Our objective was to calibrate our pipeline by estimating how many of the subscribers within certain age or race groups actually self-disclosed their age and race on Twitter and were captured by our pipeline. The estimated rates (of detection/self-disclosure) were then used as weights to normalize the age/race proportions obtained from

the NPMU dataset. The script for collecting the streaming data is included in Supplementary File, "data-collector-Random-user.py".

Calculation of the Proportions for People Who Use Substances Based on Twitter Data. For each subscriber's characteristics (gender, age, and race), we used the number of Twitter subscribers in each category, inferred by the corresponding pipelines, to estimate the proportion. For age and race, this calculation was limited to the Twitter subscribers whose age or race can be inferred. The proportions were further normalized via the rates of detection as:

$$\text{normalized proportion} = \frac{1}{\text{normalization factor}} \times \frac{(\text{raw proportion})}{(\text{rate of detection})}.$$

For example, if 10% of the Twitter subscribers are Black and only 1% of the random Twitter subscribers disclose that they are Black, then we can estimate that roughly 1 in 10 Black Twitter subscribers self-disclose their race (rate of detection). If then we captured 100 people who reported NPMU of stimulants and disclosed that they are Black, we estimate that roughly 1,000 people who use stimulants captured in our pipeline are Black and use this number to calculate the normalized proportion. For the race proportions, since AIAN, NHOPI, and Asian were combined into others category in the 2018 Twitter Survey, we only calibrated for the others category as a whole and assumed that their relative proportions are the same as obtained using the race characterization pipeline. We note that our normalization procedure was designed based on the weighting processing in the survey sampling described in Lavallée and Beaumont (72).

Calculation of Proportions for the NSDUH Data. We established the baseline for the NSDUH/US Census based on Table 12.1A (age) and Table 12.2A (gender and race) in the 2019 NSDUH. We calculated the gender, age, and race proportion through the estimated "Numbers in Thousands." For age, we used the "Total (2019)" column. For the age and gender, we used the "Age 12+ (2019)" column. For the gender and race for people who report NPMU, we again used the "Age 12+ (2019)" column on Table 1.47A (stimulants), Table 1.53A (tranquilizers), and table 1.44A (pain relievers). For age, we used the "Misuse in the past Year (2019)" column on Table 1.14A (stimulants), Table 1.16A (tranquilizers), and Table 1.13A (pain relievers).

Calculation of Proportions for the NEDS Data. We calculated the gender and age proportions of the EDVs using the "No." column for all opioid poisoning on Supplemental Table 2C in the Annual Surveillance Report of Drug-related Risks and Outcome (40). The weighted estimates provided in the table are from the NEDS 2016.

Estimation of 95% CIs. For the Twitter data and the test performance of the pipelines, the CIs are estimated via bootstrapping. For the NSDUH and NEDS data, the 95% CIs are estimated using simulation. For each category, we approximate the distribution as a normal distribution with the reported number as the mean and the SE as the SD. We then repeatedly sampled the joint distribution for all the categories in the targeted demographics and calculated the proportions, assuming each category is independent of each other. The 95% CIs were then constructed using the 0.025 and the 0.975 quantiles within the list of proportions of the given category. For the NEDS data, the SEs for estimated numbers ("No." column) were estimated through the SEs ("SE" column) for rates ("Rate" column) as $(\text{No.}) \times \frac{(\text{SE})}{(\text{Rate})}$.

Data, Materials, and Software Availability. Data (IDs of all tweets included in this study) have been deposited in Zenodo "Can accurate demographic information about people who use prescription medications non-medically be derived from Twitter?" URL: <https://doi.org/10.5281/zenodo.7401617>. The texts for the tweets, if they are still publicly available, can be obtained using the IDs via the Twitter API (73). Please see *SI Appendix* for details.

ACKNOWLEDGMENTS. Research reported in this publication was supported by National Institute on Drug Abuse of the National Institutes of Health (NIH) under award number R01DA046619 and R01DA057599. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Author affiliations: ^aDepartment of Biomedical Informatics, School of Medicine, Emory University, Atlanta, GA 30322; ^bDepartment of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN 37232; ^cDepartment of Emergency Medicine, Icahn School of Medicine at Mount Sinai, New York, NY 10029; ^dDepartment of Behavioral Sciences and Health Education, Rollins School of Public Health, Emory University, Atlanta, GA 30322; ^eDepartment of Emergency Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104; and ^fDepartment of Biomedical Engineering, Georgia Institute of Technology and Emory University, Atlanta, GA 30332

1. H. Jalal *et al.*, Changing dynamics of the drug overdose epidemic in the United States from 1979 through 2016. *Science* **361**, eaau1184 (2018).
2. C. L. Mattson *et al.*, Trends and geographic patterns in drug and synthetic opioid overdose deaths—United States, 2013–2019. *Morb. Mortal. Wkly. Rep.* **70**, 202 (2021).
3. F. B. Ahmad, J. A. Cisewski, L. M. Rossen, P. Sutton, "Provisional drug overdose death counts" (National Center for Health Statistics, Centers for Disease Control and Prevention, Hyattsville, MD, 2022).
4. Office of National Drug Control Policy, The White House, "Biden-Harris Administration Calls for Historic Levels of Funding to Prevent and Treat Addiction and Overdose" (The White House, 2021).
5. B. L. Cook, M. Alegria, Racial-ethnic disparities in substance abuse treatment: the role of criminal history and socioeconomic status. *Psychiatr. Serv.* **62**, 1273–1281 (2011).
6. A. Acevedo *et al.*, Disparities in the treatment of substance use disorders: Does where you live matter? *J. Behav. Health Serv. Res.* **45**, 533–549 (2018).
7. J. Mennis, G. J. Stahler, Racial and ethnic disparities in outpatient substance use disorder treatment episode completion for different substances. *J. Subst. Abuse Treat.* **63**, 25–33 (2016).
8. K. Burrell, C. McCuistian, J. Szapocznik, Racial/ethnic equity in substance use treatment research: The way forward. *Addict. Sci. Clin. Pract.* **16**, 1–6 (2021).
9. National Academies of Sciences, Engineering, Medicine, *Communities in Action: Pathways to Health Equity*, J. N. Weinstein, A. Geller, Y. Negussie, A. Baciu, Eds. (The National Academies Press, Washington, DC, 2017), p. 582, 10.17226/24624.
10. N. M. Astone, S. Martin, L. Aron, "Death Rates for US Women Ages 15 to 54" (Urban Institute, Washington, DC, 2015).
11. P. A. Lagisetty, R. Ross, A. Bohnert, M. Clay, D. T. Maust, Buprenorphine treatment divide by race/ethnicity and payment. *JAMA Psychiatry* **76**, 979–981 (2019).
12. H. Hansen, C. Siegel, J. Wanderling, D. DiRocco, Buprenorphine and methadone treatment for opioid dependence by income, ethnicity and race of neighborhoods in New York City. *Drug Alcohol Depend.* **164**, 14–21 (2016).
13. Substance Abuse and Mental Health Services Administration, "The opioid crisis and the Black/African American population: An urgent issue" (Publication ID PEP20-05-02-001, Office of Behavioral Health Equity, Substance Abuse and Mental Health Services Administration, Rockville, MD, 2020).
14. Substance Abuse and Mental Health Services Administration, "The opioid crisis and the Hispanic/Latino population: An urgent issue" (Publication ID PEP20-05-02-002, Office of Behavioral Health Equity, Substance Abuse and Mental Health Services Administration, Rockville, MD, 2020).
15. S. E. Hadland *et al.*, Trends in receipt of buprenorphine and naltrexone for opioid use disorder among adolescents and young adults, 2001–2014. *JAMA Pediatr.* **171**, 747–755 (2017).
16. H. Hedegaard, A. M. Miniño, M. R. Spencer, M. Warner, Drug overdose deaths in the United States, 1999–2020. *NCHS Data Brief* (2021).
17. T. Townsend *et al.*, Racial/ethnic and geographic trends in combined stimulant/opioid overdoses, 2007–2019. *Am J Epidemiol.* (2022).
18. Substance Abuse and Mental Health Services Administration, "Key substance use and mental health indicators in the United States: Results from the 2019 National Survey on Drug Use and Health - Detailed Tables" (HHS Publication No. PEP20-07-01-001, NSDUH Series H-55). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration (2020). Retrieved from <https://www.samhsa.gov/data/> (April 1, 2022).
19. U.S. Department of Health and Human Services Office of Minority Health, "Minority population profiles" (Office of Minority Health, Rockville, MD, 2021) (May 1, 2022).
20. A. Kolodny, T. R. Frieden, Ten steps the federal government should take now to reverse the opioid addiction epidemic. *JAMA* **318**, 1537–1538 (2017).
21. Centers for Disease Control and Prevention, "Data modernization initiative strategic implementation plan" (Centers for Disease Control and Prevention, Atlanta, GA, 2021). Retrieved from: <https://www.cdc.gov/surveillance/pdfs/final-dmi-implementation-strategic-plan-12-22-21.pdf> (April 1, 2022).
22. Substance Abuse and Mental Health Services Administration, "Key substance use and mental health indicators in the United States: results from the 2018 National Survey on Drug Use and Health" (HHS Publication No. PEP19-5068, NSDUH Series H-54). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration. Retrieved from <https://www.samhsa.gov/data/> (April 1, 2022).
23. D. D. Gummin *et al.*, 2019 Annual report of the American Association of poison control centers' National Poison Data System (NPDS): 37th annual report. *Clin. Toxicol.* **58**, 1360–1541 (2020).
24. S. Slavova, T. L. Bunn, J. Talbert, Drug overdose surveillance using hospital discharge data. *Public Health Rep.* **129**, 437–445 (2014).
25. S. Slavova, P. Rock, H. M. Bush, D. Quesinberry, S. L. Walsh, Signal of increased opioid overdose during COVID-19 from emergency medical services data. *Drug Alcohol Depend.* **214**, 108176 (2020).
26. C. Clifford, M. Sethi, D. Cox, A. F. Manini, First-line vasopressor and mortality rates in ED patients with acute drug overdose. *J. Med. Toxicol.* **17**, 1–9 (2021).
27. H. Hedegaard, A. M. Miniño, M. Warner, "Drug overdose deaths in the United States, 1999–2018" in *NCHS Data Brief* (Centers for Disease Control and Prevention, 2020).
28. N. D. Volkow, Collision of the COVID-19 and addiction epidemics. *Ann. Int. Med.* **173**, 61–62 (2020).
29. A. M. Lopez *et al.*, Co-use of methamphetamine and opioids among people in treatment in Oregon: A qualitative examination of interrelated structural, community, and individual-level factors. *Int. J. Drug Policy* **91**, 103098 (2021).

30. D. Ciccarone, The rise of illicit fentanyl, stimulants and the fourth wave of the opioid overdose crisis. *Curr. Opin. Psychiatry* **34**, 344–350 (2021).
31. A. Sarker *et al.*, Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from Twitter. *Drug Safety* **39**, 231–240 (2016).
32. K. O'Connor *et al.*, Pharmacovigilance on twitter? Mining tweets for adverse drug reactions. *AMIA Annu. Symp. Proc.* **2014**, 924 (2014).
33. S. Amir, M. Dredze, J. W. Ayers, "Mental health surveillance over social media with digital cohorts" in *Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology* (Association for Computational Linguistics, Minneapolis, MN, 2019), pp. 114–120.
34. D. MacLean, S. Gupta, A. Lembke, C. Manning, J. Heer, "Forum77: An analysis of an online health forum dedicated to addiction recovery" in *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (Association for Computing Machinery, Vancouver, BC, 2015), pp. 1511–1526.
35. C. L. Hanson *et al.*, Tweaking and tweeting: Exploring Twitter for nonmedical use of a psychostimulant drug (Adderall) among college students. *J. Med. Int. Res.* **15**, e62 (2013).
36. R. L. Graves *et al.*, Opioid discussion in the Twittersphere. *Subst. Use Misuse* **53**, 2132–2139 (2018).
37. M. Chary *et al.*, Epidemiology from tweets: Estimating misuse of prescription opioids in the USA from social media. *J. Med. Toxicol.* **13**, 278–286 (2017).
38. A. Sarker *et al.*, Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from Twitter. *Drug Saf* **39**, 231–240 (2016).
39. A. Sarker, G. Gonzalez-Hernandez, Y. Ruan, J. Perrone, Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. *JAMA Netw. Open* **2**, e1914672 (2019).
40. Centers for Disease Control and Prevention, "Annual surveillance report of drug-related risks and outcomes—United States Surveillance Special Report" (Centers for Disease Control and Prevention, U.S. Department of Health and Human Services, Atlanta, GA, 2019). Retrieved from <https://www.cdc.gov/drugoverdose/pdf/pubs/2019-cdc-drug-surveillance-report.pdf> (April 12, 2022).
41. M. A. Al-Garadi *et al.*, Text classification models for the automatic detection of nonmedical prescription medication use from social media. *BMC Med. Inform. Decis. Mak.* **21**, 27 (2021).
42. S. Wojcik, A. Hughes (2019) Sizing Up Twitter Users. (Pew Research Center), pp U.S. adult Twitter users are younger and more likely to be Democrats than the general public. Most users rarely tweet, but the most prolific 10% create 80% of tweets from adult U.S. users.
43. United States Census Bureau, "Census Bureau releases estimates of undercount and overcount in the 2020 census" (Press release number CB22-CN.02, United States Census Bureau, Washington, DC, 2022).
44. Substance Abuse and Mental Health Services Administration, "Key substance use and mental health indicators in the United States: Results from the 2020 National Survey on Drug Use and Health" (HHS Publication No. PEP21-07-01-003, NSDUH Series H-56). Rockville, MD: Center for Behavioral Health Statistics and Quality, Substance Abuse and Mental Health Services Administration (2021). Retrieved from <https://www.samhsa.gov/data/> (April 5, 2022).
45. M. Fendrich, T. P. Johnson, J. S. Wislar, A. Hubbell, V. Spiehler, The utility of drug testing in epidemiological research: Results from a general population survey. *Addiction* **99**, 197–208 (2004).
46. M. Fendrich, T. P. Johnson, Race/ethnicity differences in the validity of self-reported drug use: Results from a household survey. *J. Urban Health* **82**, iii67–iii81 (2005).
47. T. Johnson, M. Fendrich, Modeling sources of self-report bias in a survey of drug use epidemiology. *Ann. Epidemiol.* **15**, 381–389 (2005).
48. J. R. Bromberg *et al.*, Methodology and demographics of a brief adolescent alcohol screen validation study. *Pediatr. Emerg. Care* **35**, 737–744 (2019).
49. A. Percy, S. McAlister, K. Higgins, P. McCrystal, M. Thornton, Response consistency in young adolescents' drug use self-reports: A recanting rate analysis. *Addiction* **100**, 189–196 (2005).
50. J. D. Burger, J. Henderson, G. Kim, G. Zarrella, "Discriminating gender on Twitter" in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Edinburgh, Scotland, 2011), pp. 1301–1309.
51. W. Liu, D. Ruths, "What's in a name? using first names as features for gender inference in twitter" in *2013 AAAI Spring Symposium Series* (Association for the Advancement of Artificial Intelligence, Washington, DC, 2013).
52. M. Sap, "Developing age and gender predictive lexica over social media" in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (Association for Computational Linguistics, Doha, Qatar, 2014), pp. 1146–1151.
53. Z. Wang *et al.*, "Demographic inference and representative population estimates from multilingual social media data" in *The World Wide Web Conference* (Association for Computing Machinery, San Francisco, CA, 2019), pp. 2056–2067.
54. A. A. Morgan-Lopez, A. E. Kim, R. F. Chew, P. Ruddle, Predicting age groups of Twitter users based on language and metadata features. *PLoS One* **12**, e0183537 (2017).
55. R. Garcia-Guzman *et al.*, Trend-based categories recommendations and age-gender prediction for pinterest and twitter users. *Appl. Sci.* **10**, 5957 (2020).
56. A. Pandya, M. Oussalah, P. Monachesi, P. Kostakos, On the use of distributed semantics of tweet metadata for user age prediction. *Future Generat. Comput. Syst.* **102**, 437–452 (2020).
57. D. Preojiuc-Pietro, L. Ungar, "User-level race and ethnicity predictors from twitter text" in *Proceedings of the 27th International Conference on Computational Linguistics* (Association for Computational Linguistics, Santa Fe, NM, 2018), pp. 1534–1545.
58. J. Ye, S. Skiena, "The secret lives of names? name embeddings from social media" in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Association for Computing Machinery, Anchorage, AK, 2019), pp. 3000–3008.
59. S. Xu *et al.*, Leveraging social media to promote public health knowledge: Example of cancer awareness via Twitter. *JMIR Public Health Surveill.* **2**, e17 (2016).
60. S.-G. Jung, J. An, H. Kwak, J. Salminen, B. J. Jansen, "Assessing the accuracy of four popular face recognition tools for inferring gender, age, and race" in *Twelfth International AAAI Conference on Web and Social Media* (Association for the Advancement of Artificial Intelligence, Palo Alto, CA, 2018).
61. Y.-C. Yang, M. A. Al-Garadi, J. S. Love, J. Perrone, A. Sarker, Automatic gender detection in Twitter profiles for health-related cohort studies. *JAMIA Open* **4**, ooab042 (2021).
62. J. R. Friedman, H. Hansen, Evaluation of increases in drug overdose mortality rates in the US by race and ethnicity before and during the COVID-19 pandemic. *JAMA Psychiatry* **79**, 379–381 (2022), 10.1001/jamapsychiatry.2022.0004.
63. G. Papazisis, I. Tsakiridis, S. Sifis, Nonmedical use of prescription drugs among medical students and the relationship with illicit drug, tobacco, and alcohol use. *Subst. Abuse.* **12**, 1178221818802298 (2018).
64. Substance Abuse and Mental Health Services Administration, "Behavioral health equity" (2022).
65. National Institute of Health, "Inclusion of women and minorities as participants in research involving human subjects" (2022).
66. National Institute on Minority Health and Health Disparities, "NIH Minority health and health disparities strategic plan 2021-2025" (2021).
67. K. O'Connor, A. Sarker, J. Perrone, G. G. Hernandez, Promoting reproducible research for characterizing nonmedical use of medications through data annotation: description of a Twitter corpus and guidelines. *J. Med. Int. Res.* **22**, e15861 (2020).
68. A. Sarker, G. Gonzalez-Hernandez, An unsupervised and customizable misspelling generator for mining noisy health-related text sources. *J. Biomed. Inform.* **88**, 98–107 (2018).
69. Y. Liu *et al.*, Roberta: A robustly optimized bert pretraining approach. arXiv [Preprint] (2019), <https://arxiv.org/abs/1907.11692> (Accessed 5 March 2021).
70. S. Volkova, T. Wilson, D. Yarowsky, "Exploring demographic language variations to improve multilingual sentiment analysis in social media" in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Seattle, Washington, USA, 2013), pp. 1815–1827.
71. S. Bird, E. Loper, E. Klein, *Natural Language Processing with Python* (O'Reilly Media Inc., 2009).
72. P. Lavallée, J. Beaumont, Why we should put some weight on weights. *Survey Insights: Methods from the Field, Weighting: Practical Issues and 'How to Approach, Invited article* (2015). (SMIF-2015-00001. [CrossRef][Google Scholar]).
73. A. Sarker, Can accurate demographic information about people who use prescription medications non-medically be derived from Twitter? Zenodo. <https://zenodo.org/record/7401617#Y9dE7zMLJ8>. Deposited 5 December 2022.