Patterns of Routes of Administration and Drug Tampering for Nonmedical Opioid Consumption: Data Mining and Content Analysis of Reddit Discussions

Duilio Balsamo¹, MS; Paolo Bajardi², PhD; Alberto Salomone³, PhD; Rossano Schifanella²,4, PhD

¹Department of Mathematics, University of Turin, Turin, Italy
²ISI Foundation, Turin, Italy
³Department of Chemistry, University of Turin, Turin, Italy
⁴Department of Computer Science, University of Turin, Turin, Italy

Corresponding Author:
Duilio Balsamo, MS
Department of Mathematics
University of Turin
Via Carlo Alberto, 10
Turin, 10124
Italy
Phone: 39 3331371972
Email: duilio.balsamo@unito.it

Abstract

Background: The complex unfolding of the US opioid epidemic in the last 20 years has been the subject of a large body of medical and pharmacological research, and it has sparked a multidisciplinary discussion on how to implement interventions and policies to effectively control its impact on public health.

Objective: This study leverages Reddit, a social media platform, as the primary data source to investigate the opioid crisis. We aimed to find a large cohort of Reddit users interested in discussing the use of opioids, trace the temporal evolution of their interest, and extensively characterize patterns of the nonmedical consumption of opioids, with a focus on routes of administration and drug tampering.

Methods: We used a semiautomatic information retrieval algorithm to identify subreddits discussing nonmedical opioid consumption and developed a methodology based on word embedding to find alternative colloquial and nonmedical terms referring to opioid substances, routes of administration, and drug-tampering methods. We modeled the preferences of adoption of substances and routes of administration, estimating their prevalence and temporal unfolding. Ultimately, through the evaluation of odds ratios based on co-mentions, we measured the strength of association between opioid substances, routes of administration, and drug tampering.

Results: We identified 32 subreddits discussing nonmedical opioid usage from 2014 to 2018 and observed the evolution of interest among over 86,000 Reddit users potentially involved in firsthand opioid usage. We learned the language model of opioid consumption and provided alternative vocabularies for opioid substances, routes of administration, and drug tampering. A data-driven taxonomy of nonmedical routes of administration was proposed. We modeled the temporal evolution of interest in opioid consumption by ranking the popularity of the adoption of opioid substances and routes of administration, observing relevant trends, such as the surge in synthetic opioids like fentanyl and an increasing interest in rectal administration. In addition, we measured the strength of association between drug tampering, routes of administration, and substance consumption, finding evidence of understudied abusive behaviors, like chewing fentanyl patches and dissolving buprenorphine sublingually.

Conclusions: This work investigated some important consumption-related aspects of the opioid epidemic using Reddit data. We believe that our approach may provide a novel perspective for a more comprehensive understanding of nonmedical abuse of opioids substances and inform the prevention, treatment, and control of the public health effects.

(J Med Internet Res 2021;23(1):e21212) doi: 10.2196/21212
Background

In the last decade, the United States witnessed an unprecedented growth of deaths due to opioid drugs [1], which sparked from overprescriptions of semisynthetic opioid pain medication such as oxycodone and hydromorphone and evolved in a surge of abuse of illicit opioids like heroin [2,3] and powerful synthetic opioids like fentanyl [4,5]. Alongside traditional medical, pharmacological, and public health studies on the nonmedical adoption of prescription opioids [6-14], several phenomena related to the opioid epidemic have recently been successfully tackled through a digital epidemiology [15-18] approach. Researchers have used digital and social media data to perform various tasks, including detecting drug abuse [19,20], forecasting opioid overdose [21], studying transition into drug addiction [22], predicting opioid relapse [23], and discovering previously unknown treatments for opioid addiction [24]. A few recent studies investigated the temporal unfolding of the opioid epidemic in the United States by leveraging complementary data sources different from the official US Centers for Disease Control and Prevention data [2,25-28] and using social media like Reddit [29,30].

Pharmacology research is interested in understanding the consequences of various routes of administration (ROA), that is, the paths by which a substance is taken into the body [6,31,32], due to the different effects and potential health-related risks tied to them [10,33,34]. Researchers have estimated the prevalence of routes of administration for nonmedical prescription opioids [9,31,32,35] and opiates [36,37]; however, they rarely consider less common ROA, such as rectal, transdermal, or subcutaneous administration [32,38], leaving the mapping of nonmedical and unconventional administration behaviors greatly unexplored [39,40]. Many of these studies [31,32,35] acknowledge that drug tampering, that is, the intentional chemical or physical alteration of medications [41], is an important constituent of drug abuse. The alteration of the pharmacokinetics of opioids through drug-tampering methods, together with unconventional administration, may potentially lead to very different addictive patterns and ultimately have unexpected health-associated risks [33]. Research has also been focused on developing tamper-resistant and abuse-deterrent drug formulations. However, to the best of our knowledge, no large-scale empirical evidence has been found to unveil the relationships between substance manipulation, unconventional ROA, and nonmedical substance administration.

Goals

This paper seeks to complement current studies widening the understanding of opioid consumption patterns by using Reddit, a social content aggregation website, as the primary data source. This platform is structured into subreddits, user-generated and user-moderated communities dedicated to the discussion of specific topics (Multimedia Appendix 1). Due to fair guarantees of anonymity, no limits on the number of characters in a post, and a large variety of debated topics, this platform is often used to uninhibitedly discuss personal experiences [42]. Reddit constitutes a nonintrusive and privileged data source to study a variety of issues [43,44], including sensitive topics such as mental health [45], weight loss [46], gender issues [47], and substance abuse [22,24].

This study’s contributions are manifold. First, leveraging and expanding a recent methodology proposed by Balsamo et al [30], we identified a large cohort of opioid firsthand users (ie, Reddit users showing explicit interest in firsthand opioid consumption) and characterized their habits of substance use, administration, and drug tampering over a period of 5 years. Second, using word embeddings, we identified and cataloged a large set of terms describing practices of nonmedical opioids consumption. These terms are invaluable to performing exhaustive and at-scale analyses of user-generated content from social media, as they include colloquialisms, slang, and nonmedical terminology that is established on digital platforms and hardly used in the medical literature. We provided a longitudinal perspective on online interest in the opioids discourse and a quantitative characterization of the adoption of different ROA, with a focus on the less-studied yet emerging and relevant practices. We have made available the ROA taxonomy and the corresponding vocabulary to the research community. Third, we quantified the strength of association between ROA and drug-tampering methods to better characterize emerging practices. Finally, we investigated the interplay between the previous 3 dimensions, measuring odds ratios to shed light on the “how” and “what” facets of the opioid consumption phenomenon. We studied a wide spectrum of opioid forms, referred to as “opioids” throughout, ranging from prescription opioids to opiates and illegal opioid formulations. To the best of our knowledge, our contributions are original in both breadth and depth, outlining a detailed picture of nonmedical practices and abusive behaviors of opioid consumption through the lenses of digital data.

Methods

Data

We refer to a publicly available Reddit data set [48] that contains all the subreddits published on the platform since 2007 [44,49]. In this work, we analyzed the textual part of the submissions and the comments collected from 2014 to 2018. We preprocessed each year separately, filtering out the subreddits with less than 100 comments in a year. We used spaCy [50] to remove English stop words, inflectional endings, and tokens with less than 100 yearly appearances. We adopted a bag-of-words model, resulting in a vocabulary of different lemmas for each year. Vocabulary sizes ranged from 300,000 to 700,000 lemmas, with a size growth of approximately 30% each year. In Table 1, the number of unique comments and unique active users per year is reported. A steady growth of...
approximately 30% per year both in the volume of conversations and in the active user base is observed.

All the analyses in this work were performed on a subset of subreddits related to opioid consumption, which were identified using the procedure described here. For space constraints, we restricted the analyses of odds ratios to comments and submissions created during 2018. Similar to a vast body of

users’ activities on social media platforms [51-53], the distribution of posts per user shows a heavy tail, with the majority of users posting few comments and the remaining minority (eg, core users and subreddit moderators) producing a large portion of the content. Moreover, a nonnegligible percentage of posts, respectively 25% and 7% of submissions and comments, were produced by authors who deleted their usernames.

Table 1. Data set statistics.

<table>
<thead>
<tr>
<th>Year</th>
<th>Reddit comments, n</th>
<th>Reddit authors, n</th>
<th>Opiates subreddits, n</th>
<th>Opiates comments, n</th>
<th>Opiates authors, n</th>
<th>Authors’ prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td>2014</td>
<td>545,720,071</td>
<td>8,149,234</td>
<td>19</td>
<td>386,984</td>
<td>12,381</td>
<td>0.0015</td>
</tr>
<tr>
<td>2015</td>
<td>699,245,245</td>
<td>10,673,990</td>
<td>19</td>
<td>470,609</td>
<td>15,888</td>
<td>0.0015</td>
</tr>
<tr>
<td>2016</td>
<td>840,575,089</td>
<td>12,849,603</td>
<td>25</td>
<td>612,619</td>
<td>21,791</td>
<td>0.0017</td>
</tr>
<tr>
<td>2017</td>
<td>1,045,425,499</td>
<td>14,219,062</td>
<td>30</td>
<td>866,023</td>
<td>28,358</td>
<td>0.0020</td>
</tr>
<tr>
<td>2018</td>
<td>1,307,123,219</td>
<td>18,158,464</td>
<td>25</td>
<td>919,036</td>
<td>33,700</td>
<td>0.0019</td>
</tr>
</tbody>
</table>

Analytical Pipeline

The methodology adopted in this paper consists of several steps. First, we identified a cohort of opioid firsthand users and the subreddits related to opioid consumption through a semiautomatic algorithm. Second, we trained a word-embedding language model to capture the latent semantic features of the discourse on the nonmedical use of opioids. Third, we exploited the embedded vectors to extend an initial set of medical terms known from the literature, (eg, opioid substance names, ROA, and drug-tampering methods) to nonmedical and colloquial expressions. The terms were organized in a taxonomy that provides a conceptual map on the topic. Moreover, we studied the temporal evolution of the popularity of the main opioid substances and ROA. Ultimately, we measured the strength of the associations between opioid substances, routes of administration, and drug-tampering techniques in 2018.

Identification of Firsthand Opioid Consumption on Reddit

We leveraged a semiautomatic information retrieval algorithm developed to identify relevant content related to a topic of interest [30] to collect opioid-related conversations on Reddit yearly. This approach aims at retrieving topic-specific documents by expressing a set of initial keywords of interest; here, it identified relevant subspaces of discussion via an iterative query expansion process, retaining a list of terms Q, and a list of subreddits S, ranked by relevance for each year.

We merged all the query terms in a set $Q = \cup_{i=1}^{k} Q_i$ containing 67 terms. To ensure that the sets $S_i$ were effectively referring to the opioid-related topics and in particular to nonmedical opioid consumption, we performed a manual inspection on the union

$T = \cup_{i=1}^{k} S_i$ of the top 150 subreddits for each year, for a total of 554 subreddits. Three independent annotators, including a domain expert specialized in antidoping analyses, read a random sample of 30 posts, checking for subreddits (1) mostly focused on discussing the use of opioids, (2) mostly focused on firsthand usage, and (3) not focused on medical treatments. This yielded a total of 32 selected subreddits, with a Fleiss $\kappa$ intrarater agreement of $\kappa=0.731$, which suggests a substantial agreement, according to Landis and Koch [54]. Multimedia Appendix 2 presents a complete list of the subreddits broken down by year.

Automatic language detection, performed with langdetect [55], cld2 [56], and cld3 [57], showed that the majority of posts (about 90%) were in English, approximately 5% were non-English messages, and the rest were too short or full of jargon and emojis to algorithmically detect any language. Assuming that an author who writes in one of the selected subreddits is personally interested in the topic, we identified a cohort of 86,445 unique opioid firsthand users involved in direct discussions of opioid usage across the period of study. Summary statistics are reported in Table 1. In particular, for each year, we computed the number of unique active users and the volume of comments shared, as well as the user’s relative prevalence over the entire amount of Reddit activity. We observed growth from 2014 to 2017, ranging from 15 to 19 users interested in opioid consumption out of every 100,000 Reddit users.

Vocabulary Expansion

The methodology to extend the vocabulary on opioid-related domains with user-generated slang and colloquial forms was implemented in 2 steps. First, we trained a word-embedding model (word2vec [58]), which learns semantic relationships in the corpus during training and maps their terms to vectors in a latent vectorial space, with all the comments and submissions in our subreddit data set (relevant training parameters are displayed in Multimedia Appendix 3). Second, starting from a set of seed terms $K$ (eg, a list of known opioid substances), we expanded the vocabulary by navigating the semantic neighborhood $E^w_n = \text{neighbors}(w, n)$ of each element $w \in K$ in the embedded space, considering the $n=20$ semantically closest elements in terms of cosine similarity. We merged the results in a candidate expansion set $E = \cup_{w \in K} E^w_n$, together with the seed terms $K$ if not already included. Based on the knowledge of a domain expert (a clinical and forensic toxicologist) and with the help of search engine queries and a crowdsourced online dictionary for slang words and phrases (Urban Dictionary [59]) to understand the most unusual terms, we manually selected
and categorized the relevant neighboring terms, obtaining an extended vocabulary $V$. Figure 1 shows an example of the expansion procedure in which the high-dimensional vectors are projected to 2 dimensions using the uniform manifold approximation and projection (UMAP) algorithm [60].

As a sensitivity analysis, we compared the effectiveness of an alternative embedding model (GloVe [61]) for topical coherence. In the case of vocabulary expansion of opioid substance terms, that is, using $K = Q$ as seeds, the 2 models captured 100 terms in common out of their respective candidate terms, with word2vec showing a higher number and a larger percentage of accepted terms (Table 2). Moreover, the volume of comments that included an accepted term was almost double when using the vocabulary of word2vec rather than the vocabulary of GloVe. Hence, we chose word2vec as the reference word-embedding model.

Figure 1. Two-dimensional projection of the word2vec embedding, modeling the semantic relationships among terms in the Reddit opioids data set. Filled markers represent the seed terms $K$. Expansion terms, represented with hollow markers, are colored according to their respective initial term if accepted or in gray if discarded. The nature of the relationships between neighboring terms varies, representing (1) equivalence (eg, synonyms), (2) common practices (eg, the use of methadone for addiction maintenance), or (3) co-use (eg, the cluster of heroin, cocaine, and methamphetamine).
Table 2. Comparison of term expansions of opioid substances for the 2 trained models.

<table>
<thead>
<tr>
<th>Model</th>
<th>Candidate terms, n</th>
<th>Accepted terms, n (%)</th>
<th>Comments², n</th>
</tr>
</thead>
<tbody>
<tr>
<td>word2vec</td>
<td>297</td>
<td>128 (43.1)</td>
<td>225,165</td>
</tr>
<tr>
<td>GloVe</td>
<td>369</td>
<td>110 (29.8)</td>
<td>144,564</td>
</tr>
</tbody>
</table>

²Comments in the corpus mentioning at least one term of the respective accepted terms for vocabulary expansion.

Strength of Association Between Opioid Substances, ROA, and Drug Tampering

We evaluated the odds ratios (ORs) to quantify the pairwise strength of the association between substance use and ROA, substance use and drug-tampering methods, and ROA and drug-tampering methods. Under the assumption that co-mention was a proxy for associating a substance to its ROA (or drug-tampering method), we focused on the posts that contained a reference to terms in each domain, evaluating contingency tables and odds ratios. Odds ratios, significance, and confidence intervals were estimated using chi-square tests implemented in the statsmodel Python package [62], with the significance level set to $\alpha=.01$. As a sensitivity analysis, we assessed the effect of the proximity of terms on the characterization of odds ratios. We modified the definition of co-occurrence, introducing a distance threshold $\rho$ at sentence level. We explored the range $\rho \in \{0...5\}$, where $\rho=0$ indicates that co-occurrence appears within the same sentence, and $\rho>0$ measures the distance in both directions (eg, $\rho=1$ for the preceding and consecutive sentences). The value $\rho=\infty$ indicates the scenario in which we considered the entire post as reference. Accordingly, given a threshold $\rho$ in the construction of the contingency table, the co-occurrence event between 2 terms is conditioned to their distance being less than or equal to $\rho$. Conversely, we considered terms to be separate events in cases of distance above the threshold. It is important to consider that the OR measures do not imply any form of causation but rather surface correlations that could be used in hypothesis formation. To better interpret the results of this analysis, in some cases, manual inspection of the comments mentioning the variables under investigation was performed following the directives on privacy and ethics (see the “Ethics and Privacy” section).

Results

Characterizing Interest in Opioids, ROA, and Drug-Tampering Methods

We applied the methodology described in the “Vocabulary Expansion” section to extract and expand domain-specific vocabularies and to characterize the temporal unfolding of interest in different opioid substances, routes of administration, and drug-tampering methodologies. We started from a review of the relevant medical research, collecting an initial set of terms referring to the most common opioid substances, ROA [6,10,31,34,38,39,41,63,64], and drug-tampering methods [41,63]. We expanded the original set with neighboring terms in a low-dimensional embedding space, and the outputs were reviewed and organized by a domain expert. The resulting vocabulary for opioid substances is shown in Table 3. It is worth noting that the vocabulary expansion procedure considerably increased the richness of the terminology related to the domain of interest and, consequently, the volume of conversations on Reddit that contained these terms. For example, for the heroin category, we observed a 62% growth in the retrieved relevant conversations (Table 3). We investigated the temporal unfolding of the popularity of the opioid substances, measured as the fraction of authors mentioning a substance over the entire opioid firsthand user base, for each trimester from 2014 to 2018. A binary characterization of the mentioning behavior at the user level was considered to discount potential biases due to users with high activity. We also provided a relative measure of popularity to account for the constantly increasing volume of active users on Reddit. Figure 2 shows a decrease in the usage of heroin and a rise in fentanyl and codeine.

The resulting vocabulary for routes of administration was further organized in a 2-level hierarchical structure, reported in Table 4. It is worth noting that the taxonomy does not have a strict medical interpretation, nor was it intended to be a comprehensive review. However, it can give structure to otherwise unstructured collections of words and help in the interpretation of the results.
Table 3. Vocabulary of opioid substances. Starting from a candidate expansion set $E$, comprising 297 unique terms, the final expansion terms considered equivalent to a substance were gathered in the same class.

<table>
<thead>
<tr>
<th>Substance</th>
<th>Terms</th>
<th>$\Delta$ volume, %$^a$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heroin</td>
<td>btb, diacetyl morphine, diamorphine, dope, ecp, goofball, goofballs, gunpowder, h, heroin, heroin, heroin, smack, speedball, speedballing, speedballs, tar</td>
<td>62</td>
</tr>
<tr>
<td>Buprenorphine</td>
<td>bup, bupe, buprenorphine, butrans, sub, suboxone, subutex, zub, zubsolv</td>
<td>61</td>
</tr>
<tr>
<td>Hydrocodone</td>
<td>hydro, hydrocodone, hydrocodones, lortab, lortabs, norco, norcos, tuss, tussionex, vic, vicoden, vicodin, vicodins, vicoprofen, vics, vikes, viks, zohydro</td>
<td>38</td>
</tr>
<tr>
<td>Codeine</td>
<td>cocodamol, codein, codeine, codine, dhc, dihydrocodeine, prometh, sizzurrp, syrup</td>
<td>28</td>
</tr>
<tr>
<td>Oxymorphone</td>
<td>g74, opana, opanas, oxymorphone, panda</td>
<td>25</td>
</tr>
<tr>
<td>Tramadol</td>
<td>desmethyltramadol, dsmt, tram, tramadol, ultram</td>
<td>22</td>
</tr>
<tr>
<td>Hydromorphone</td>
<td>dil, dilauded, dilaudid, dilaudids, dillys, dily, dillys, diluadid, hydromorph, hydromorphone</td>
<td>21</td>
</tr>
<tr>
<td>Oxycodone</td>
<td>15s, 30s, codone, contin, ms, oc, ocs, oxy, oxycodone, oxycontin, oxycontins, oxycotin, oxys, perc, percoset, percocet, percocets, perc, perk, roxie, roxie, roxies, roxies, roxy, roxy</td>
<td>14</td>
</tr>
<tr>
<td>Morphine</td>
<td>kadian, morph, morphine</td>
<td>5</td>
</tr>
<tr>
<td>Fentanyl</td>
<td>acetylfentanyl, butyr, butyrfentanyl, carf, carfent, carfentanil, carfentanyl, duragesic, fent, fentanyl, fents, fentanyl, furanyl, sufentanil, u47700</td>
<td>4</td>
</tr>
<tr>
<td>Antagonist</td>
<td>nalaxone, naloxone, naltrexone, narcan, narcon, revia, viv, vivitrol</td>
<td>1</td>
</tr>
<tr>
<td>Methadone</td>
<td>mdone, methadone, methadone</td>
<td>1</td>
</tr>
</tbody>
</table>

$^a$ The increase in the volume of occurrences of a substance using the terms in the expanded vocabulary compared with only using the terms in $Q$.

$^b$ Term in $Q$.

Figure 2. Popularity of opioid substances among opioid firsthand users on Reddit. Each line represents the share of opioid firsthand users mentioning an opioid substance, measured quarterly from 2014 to 2018.
Table 4. Taxonomy defining the ROA categories and their corresponding terms. Primary ROA include all the expansion terms considered for the appropriate secondary ROA (original candidate expansion set comprised 199 unique terms).

<table>
<thead>
<tr>
<th>Primary and secondary ROA</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ingestion</strong></td>
<td></td>
</tr>
<tr>
<td>Oral</td>
<td>bolus, buccal, gulp, mouth, mouthful, oral, orally, swallow</td>
</tr>
<tr>
<td>Sublingual</td>
<td>sublingual, sublingually, tongue</td>
</tr>
<tr>
<td>Drink</td>
<td>chug, drink, pour, pourin, sip, sipper, sippin, swig, swish</td>
</tr>
<tr>
<td>Chew</td>
<td>chew, chewy, chomp, gum</td>
</tr>
<tr>
<td>General ingestion</td>
<td>ingest, ingestion</td>
</tr>
<tr>
<td><strong>Inhalation</strong></td>
<td></td>
</tr>
<tr>
<td>Intranasal</td>
<td>intranasal, intranasally, nasal, nasally, nose, nostril, rail, sniff, sniffer, sniffin, snoot, snooter, snort, snorter, tooter</td>
</tr>
<tr>
<td>General inhalation</td>
<td>breath, breathe, dab, exhale, inhalation, inhale, insufflate, insufflated, insufflating, insufflation, puff, toke, tokes, vap, vape, vaped, vapes, vaping, vapor, vaporise, vaporize, vaporizer, vapour</td>
</tr>
<tr>
<td><strong>Smoking</strong></td>
<td>bong, fume, hookah, pipe, smoke, smoker, smokin, spliff</td>
</tr>
<tr>
<td><strong>Injection</strong></td>
<td></td>
</tr>
<tr>
<td>Intramuscular</td>
<td>deltoid, imed, iming, intramuscular, intramuscularly</td>
</tr>
<tr>
<td>Subcutaneous</td>
<td>subcutaneous, subcutaneously, subq</td>
</tr>
<tr>
<td>Intravenous</td>
<td>arterial, bloodstream, intravenous, intravenously, iv, ivd, ived, ivs, vein, venous</td>
</tr>
<tr>
<td>General injection</td>
<td>bang, inject, injectable, injection, parenteral, shoot, shot</td>
</tr>
<tr>
<td>Rectally</td>
<td>anal, anally, boof, boofed, boofing, bunghole, butt, pooper, rectal, rectally</td>
</tr>
<tr>
<td><strong>Other ROA</strong></td>
<td></td>
</tr>
<tr>
<td>Dermal</td>
<td>cutaneous, dermis, transdermal, transdermally</td>
</tr>
<tr>
<td>Urogenital</td>
<td>vaginal</td>
</tr>
<tr>
<td>Intrathecal</td>
<td>intrathecal</td>
</tr>
</tbody>
</table>

aROA: routes of administration.
bSeed term K.

Figure 3 shows the estimated temporal evolution of the relative popularity of the routes of administration from 2014 to 2018, measured in quarterly snapshots. Finally, we extracted and organized the vocabulary related to drug-tampering techniques, as shown in Table 5. In this paper, we considered the act of chewing pills a second-level route of administration under the ingestion category [8,31,32] instead of a drug-tampering method, as some research might suggest [41].
Figure 3. Popularity of routes of administration among opioid firsthand users on Reddit. Each line represents the fraction of opioid firsthand users mentioning an ROA-related term, measured quarterly from 2014 to 2018. Thick lines represent the share of authors mentioning primary ROA, evaluated by aggregating the contributions of all the corresponding secondary ROA. ROA: routes of administration.

Table 5. Vocabulary of drug-tampering methods. Expansion terms referring to the same drug-tampering method are grouped in the corresponding transformation classes (original candidate expansion set comprised 179 unique terms).

<table>
<thead>
<tr>
<th>Transformation</th>
<th>Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brew</td>
<td>brew⁴, brewer, homebrew⁴</td>
</tr>
<tr>
<td>Concentrate</td>
<td>concentrate⁴, concentrate, concentration, purify</td>
</tr>
<tr>
<td>Dissolve</td>
<td>desolve, dilute, dissolve, dissolved, dissolves, dissolve⁴, solute, solution, soluble, soluable</td>
</tr>
<tr>
<td>Evaporate</td>
<td>evap, evaporate</td>
</tr>
<tr>
<td>Extract</td>
<td>cwe⁴, extract⁴, extract, extraction</td>
</tr>
<tr>
<td>Grind</td>
<td>chop, crush⁴, crushable, crusher, grind⁴, grinded, grinder, ground, pulverize</td>
</tr>
<tr>
<td>Heat</td>
<td>boil, heat⁴, melt, microwave, overheat, simmer</td>
</tr>
<tr>
<td>Infusion</td>
<td>infuse, infusion⁴, tea, tincture</td>
</tr>
<tr>
<td>Peel</td>
<td>peal, peel, shave</td>
</tr>
<tr>
<td>Soak</td>
<td>soak⁴, submerge</td>
</tr>
<tr>
<td>Wash</td>
<td>rewash, rinse, wash</td>
</tr>
</tbody>
</table>

⁴Seed term K.

Characterizing the Associations Between Opioid Substances, ROA, and Drug Tampering

To investigate the strength of association between routes of administration, drug tampering, and opioid substances and to shed light on the interplay between the “how” and the “what” dimensions of opioid consumption, we estimated the ORs, 95% confidence intervals, P values, and volume of the co-mentions among substances, routes of administration, and drug-tampering methods. The number of sentences in Reddit posts vary greatly, but the posts are generally quite short (approximately 50% of them have 2 sentences or less, as seen in Multimedia Appendix 4). However, as about 20% of posts have more than 10 sentences, one should be cautious in adopting a bag-of-words
approach to measure co-occurring terms. To limit the chance
of including spurious correlations due to the co-occurrences
of terms far apart in the posts, we conservatively selected \( \rho = 1 \) (ie,
considering only the co-occurrence of terms within a sentence
or in the first adjacent sentences) for computing the OR. Figure
4 shows in blue the results of the analysis at \( \rho = 1 \), matching 4
of the main widespread substances (ie, heroin, buprenorphine,
oxycodone, and fentanyl) with the secondary ROA (upper panel)
and the drug-tampering techniques (lower panel). Figure 5 shows
the odds ratios of primary ROA and drug-tampering methods.
For reference, the green markers represent the ORs obtained at
\( \rho = 0 \) and \( \rho = \infty \) for the same categories. Multimedia Appendices
3-5 provide the complete set of results for all the substances
identified and the secondary ROA. Due to the low
representativeness of intrathecal and urogenital ROA with most
of the tampering-related terms, we omitted those categories
from the analysis. In the plots, the associations that are not
statistically significant (\( P > .01 \)) are reported in gray, and the
horizontal lines indicate the OR and the 95% confidence
interval. The radius of the circle is proportional to the sample
of co-mentions and the dashed vertical line corresponds to an
OR of 1 for reference.

**Figure 4.** Odds ratios of the most widespread opioid substances with routes of administration (top row) and drug-tampering methods (bottom row). Labels on the right axis report the confidence interval at \( \rho = 1 \). OR: odds ratio.
**Discussion**

**Opioid Interest on Reddit**

In this work, we identified over 3 million comments on 32 subreddits focused on discussing practices and implications of firsthand opioid use. We also selected a cohort of over 86,000 Reddit users interested in this topic. Such a large data set allowed us to assess the magnitude of the online interest in opioids and model its evolution during the 5 years of study, sadly verifying its rapidly increasing popularity. By the end of 2018, the opioid epidemic remained an escalating public health threat, and at the time of writing, the opioid crisis is still calling for countermeasures at scale. Hence, we believe our large data set may constitute a valid alternative source to advise decision making and a valuable starting point for future infodemiology research.

**Vocabulary Expansion**

By observing the vocabularies in Tables 3-5 resulting from the expansion algorithm, we can ascertain the importance of enriching domain expertise with user-generated content and observe that some common features are captured across categories. Our method was able to detect synonyms and common short names, very specific acronyms (eg, “cwe” for cold water extraction [65]), slang expressions like “sippin” (often used when referring to the act of drinking codeine mixtures [63]), nicknames (eg, “panda” for oxymorphone), and polypharmacy instances (eg, “speedball” and “goofball” [66]). The vocabulary expansion underlines the use of prescription dosages (usually stamped on the tablets) in place of the commercial names of the substances (eg, “30s” for oxycodone). Moreover, we deduced that opioid firsthand users discussed variants of the substances (eg, “bth” and “ecp” for black tar heroin and East Coast powder), research chemical equivalents (eg, “u47700” [67]), and formulations intended for veterinary use (eg, sufentanil, carfentanil).

ROA vocabulary included and categorized both medical terms, adding terms scarcely considered in previous studies, like “vaping,” and nonmedical or unconventional administration terms, such as “chewing,” “snorting,” “smoking,” and “booﬁng” [39]. Our taxonomy also enabled us to disambiguate common primary ROA, such as injection and ingestion, into specific secondary ones, like subcutaneous [39] and sublingual administrations.

Finally, the drug-tampering vocabulary captured tampering methods that modify the physical status of the substances, like crushing and peeling, and some methods aiming at altering the chemical characteristics of the substances, like dissolving, washing, and heating [41]. We believe that even if this vocabulary might not be exhaustive of all drug-tampering methods, it offers a novel evidence-based perspective on the topic compared with the existing literature. The expanded vocabularies proved essential to fully incorporating the language complexity of online discussions and taboo behaviors [68] into at-scale analyses. Hopefully, our contribution might be useful in the future to find and understand new abusive behaviors that are discussed online, ultimately driving future research to yield more effective prevention methods.

**Adoption Popularity of Opioid Substances and ROA**

Considering the share of users mentioning a term to be a proxy of firsthand involvement in opioid-related activities and including topic-specific terminology, the longitudinal views in Figures 2 and 3 can be used to rank the popularity of nonmedical usage of opioid substances and ROA and their adoption trends. Ranking the substances by average share, we can see that heroin is by far the most popular substance, mentioned on average by 1 in every 3 users. Its share of users, though, is steadily
decreasing, with a loss of 10% reported in state-specific findings by Rosenblum et al [27]. Buprenorphine and oxycodone were the most mentioned prescription opioids; they showed fairly static behavior, while hydrocodone importance decreased over time [28], possibly due to more stringent prescription regulation starting in 2014 [69]. Fentanyl showed the most abrupt behavior, dramatically increasing since 2016. Its volume of mentions in 2018 increased by almost 1.5 times compared with 2014, confirming it as one of the most recent threats [5,28]. In contrast, we did not find evidence of drastic changes in oxymophine adoption after its partial ban in 2017 [70]. ROA adoption was led by injection and inhalation, which were the most popular ROA across the years, mentioned by 1 of every 3 authors at their peak. These were followed closely by ingestion. Rectal use and other ROA involved, on average, a significantly lower share of users, around 5% and less than 1%, respectively. Nevertheless, rectal administration has shown a sharp increase in popularity since 2016, almost doubling its share. Administration through inhalation was equally staggered by the intranasal and smoking categories of secondary ROA, strong indicators that this route of administration is indeed capturing nonmedical use of opioids.

This work on understanding which substances are currently gaining popularity may give prevention programs a strategic advantage, especially if consumption trends can be localized geographically [12,30,71], focusing the interventions needed to prevent early adoption of emerging dangerous substances like fentanyl. Moreover, tracking the evolution of interest in prescription opioids might be useful for evaluating the efficacy of ban policies, as in the case of oxymorphine. Understanding which ROA are the most adopted might eventually help address targeted campaigns informing users on safer practices, develop better tamper-resistant prescription drugs, and ultimately better inform the health system of the health risks specific to opioid adoption.

Characterizing the Association Between Substance Consumption, ROA, and Drug-Tampering Methods

By jointly considering the results of the odds ratios in Figures 4 and 5 and Multimedia Appendices 5-7, we can outline complex preferences for the nonmedical use of opioids, triangulating substance use, ROA, and drug-tampering methods. We noticed that the majority of substances exhibited more than one high odds ratio, both with ROA and drug-tampering methods, meaning that such substances might be consumed by users in multiple nonexclusive ways. Our analysis shows that for the most part, the expected medical and nonmedical routes of administration of each substance (ie, intended ROA or known abusive administration) had high odds ratios. For prescription opioids, oral (medical) use was often confirmed (eg, oxycodone: OR 3.6, 95% CI 3.4-3.8), while intranasal administration was often the preferred nonmedical ROA, followed by injection, especially intravenous administration (eg, hydromorphone: OR 9.1, 95% CI 8.6-9.8) [32,72]. As expected, heroin appeared to be most likely consumed through injection (OR 3.3, 95% CI 3.2-3.4) or smoking, if heated up on aluminum foil (OR 3.1, 95% CI 3.0-3.2). Heroin was the only substance that showed high correlations with this administration route. It was also reported to be snorted [64].

Besides confirming and quantifying some known behaviors, our analysis can provide additional insights on the nonmedical use of intended routes of administration. In accordance with the literature [31,32,40,73], we found evidence that abuse of prescription opioids may be associated with chewing the pills (eg, oxycodone: OR 2.7, 95% CI 2.4-3.0). From the analysis of ROA and drug-tampering methods, it appears that nonmedical oral administration was correlated with dissolving (OR 9.7, 95% CI 9.0-10.4), grinding, and washing the substances. In some cases, oral and chewing-related misuse of prescription opioids simply consisted of peeling (OR 5.1, 95% CI 2.6-9.9) the external coating, which is usually hard to chew or responsible for the extended-release effect. Even though some formulations, such as Opana ER (oxymorphine hydrochloride extended-release tablets; Endo Pharmaceuticals), are known to be tamper resistant to crushing, users can peel the tablets to get rid of the extended release coating for higher recreational effects. Injection usually requires that the substance be dissolved (OR 3.5, 95% CI 3.2-3.7), while inhalation requires that the substance be ground to powder, especially for intranasal abuse (OR 6.7, 95% CI 6.3-7.1).

Our method ultimately found evidence of unconventional nonmedical administration for most of the substances. We found a high correlation between dissolving and intranasal administration (OR 4.1, 95% CI 3.8-4.4), which may indicate the adoption of “monkey water,” the practice of dissolving soluble substances, like tar heroin and fentanyl patches, and using the resulting liquid as a nasal spray [36]. Fentanyl patches were also consumed in other unforeseen ways; an unexpectedly high OR of fentanyl and chewing (OR 2.6, 95% CI 2.2-3.0) suggests that prescription patches intended for transdermal use may be chewed for nonmedical use. Our analyses revealed the high odds of abuse of codeine via drinking (OR 4.0, 95% CI 3.7-4.3) codeine syrup, made by extracting or brewing the cough suppressants (OR 14.1, 95% CI 11.5-17.2) and forming the so-called lean or purple drank [7,63,74].

Buprenorphine, usually administered sublingually in its formulations without an antagonist, such as Subutex (buprenorphine; Indivior), and orally in combination with naloxone in the form of pills, such as Suboxone (buprenorphine-naloxone; Indivior) and Zubsolv (buprenorphine-naloxone; Orexo), measured exceptionally high odds of sublingual administration (OR 7.6, 95% CI 7.0-8.2). Evidence of nonmedical use of buprenorphine was also found in the association between dissolving and sublingual use (OR 18.9, 95% CI 16.8-21.3). Opioid firsthand users know that the opioid antagonist in buprenorphine-naloxone compounds has low bioavailability if dissolved under the tongue; hence, to achieve higher opioid effects and eliminate the antagonist, these compounds are generally taken sublingually and not through other ROA, with which buprenorphine shows negative associations.

Finally, our study shows that rectal administration is a viable and unforeseen option for the nonmedical use of some opioids, resulting in higher recreational effects, especially with hydromorphone (OR 5.2, 95% CI 4.6-6.0), morphine, and oxycodone. Rectal administration showed high correlations with the dissolving, grinding, and soaking drug-tampering methods.
methods, possible indicators of an unconventional route of administration, largely overlooked, which involves dissolving the substances in liquid water or alcohol (ie, “butt-chugging”) [39,75]. Subcutaneous administration was only weakly associated with morphine, suggesting that the practice of “skin popping” [38], which consists of injecting the substance in the tissues under the skin, is potentially not widespread.

The complex interactions between substance use, routes of administration, and drug tampering can be unified with our methodology provide a broad yet detailed perspective on the nonmedical use of opioids, evidencing abusive behaviors in which unconventional ROA and drug tampering play a key role. Knowledge about abusive behaviors could be taken into consideration by physicians during treatment programs, allowing them to favor opioid medications that are less likely to be transformed and abused. Our results should be addressed with effective health policies, driving future clinical research to better focus its efforts on understanding health-related risks and guiding the production of new tamper-resistant and abuse-deterrent opioid formulations.

Limitations and Future Work
We acknowledge some limitations in the present research. The population sampled on Reddit might have intrinsic social media biases, and it is likely not representative of the general population (eg, for gender, age, or ethnicity). Moreover, since we enrolled the users in our cohort based on their engagement in subcommunities focusing on firsthand use of opioids, we cannot exclude the possibility that in some cases, such users might have been reporting secondhand experiences, disseminating general news, or discussing intended medical drug use for pain management. We must also consider that the selected individuals were not clinically diagnosed with opioid use disorder. Future work will be devoted to building a classifier at the user level to identify individuals with opioid use disorder.

We are aware that Reddit data have some gaps [76], but since the incompleteness mostly affects the years before 2010, we consider the overall results of our work to not be significantly biased. Other limitations are related to the analytic pipeline, where we narrowed our text analysis to term counts and co-occurrences, which might have produced spillover effects in comments discussing multiple topics and could have amplified the strength of cross-associations. Future work will include n-grams and more context-based language models. Finally, it is worth stressing that the measure of association through odds ratios should not be intended by any means as an indication of causal effects. This work is an observational study focusing on the characterization of a complex and faceted social phenomenon rather than the identification of determinants or interventions, and it shares the strengths and limitations of correlational studies, especially in medical research.

Ethics and Privacy

Given the sensitive nature of the information shared, including users’ vulnerabilities and personal information, privacy and ethical considerations are paramount. In this work, we followed the guidelines and directives in Eysenbach and Till [77], which describe recommendations to ethically conduct medical research with user-generated online data, and we relied on the vast experience of research works dealing with sensitive data gathered on social media [47,78-81]. The researchers had no interactions with the users and have no interest in harming any, and the analyses were performed and reported in the spirit of knowledge, prevention, and harm reduction. In this direction, it is worth noting that the subreddits under study are of public domain, are not password protected, and have thousands of active subscribers; users were fully aware of the public nature of the content they posted and of its free accessibility on the web. Moreover, Reddit offers pseudonymous accounts and strong privacy protection, making it unlikely that the true identity of a user can be recovered. Nevertheless, in order to further protect the privacy and anonymity of the users in our data set, all information about the names of the authors was anonymized before using the data for analysis. Moreover, every analysis performed was intended to provide aggregated estimates aimed at research purposes, and this work did not include any quotes or information that focused on single authors. Following the directives in Eysenbach and Till [77], our research did not require informed consent.

Conclusions

In this work, we characterized opioid-related discussions on Reddit over 5 years, involving more than 86,000 unique users, and focused on firsthand experiences and nonmedical use. To address the complexity of the language in social media communications, especially in the presence of taboo behaviors such as drug use, we gathered a large set of colloquial and nonmedical terms that covered most opioid substances, routes of administration, and drug-tampering methods. We were able to characterize the temporal evolution of the discourse and identify notable trends, such as the surge in the popularity of fentanyl and the decrease in the relative interest in heroin. Focusing on routes of administration, we extended pharmacological and medical research with an in-depth characterization of how opioids substances are administered, since different practices imply different effects and potential health-related risks. We proposed a 2-layer taxonomy and corresponding vocabulary that enabled us to study both medical and recreational routes of administration. We demonstrated the presence of conventional nonmedical ROA (eg, intranasal administration and intravenous injection) and the spread of less conventional practices (eg, an increasing trend in rectal use). In particular, with reference to nonconventional ROA, we characterized for the first time at scale the phenomenon of drug tampering, which could have an impact on health outcomes, since it alters the pharmacokinetics of medications. The interplay between these dimensions was systematically characterized by quantitatively measuring the odds ratios, providing an insightful picture of the complex phenomenon of opioid consumption as discussed on Reddit.
Acknowledgments

PB acknowledges support from the Intesa Sanpaolo Innovation Center. The funder had no role in the study design, data collection, analysis, decision to publish, or preparation of the manuscript. RS was partially supported by the project Countering Online Hate Speech Through Effective On-line Monitoring, funded by Compagnia di San Paolo.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Schematic representation of the structure of Reddit. Reddit's most common access point is the front page, where the most relevant content of the moment is collected. The users can post on already-existing subreddits or they can create and moderate new ones on any topic of choice. In a subreddit, users can either create a new thread via a submission or indefinitely expand the conversation tree by commenting on an existing thread. The level of content moderation in a subreddit is solely decided by its moderators.

[PNG File, 71 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Subreddits discussing firsthand nonmedical use of opioids. An X marks the presence of a subreddit in a specific year.

[XLSX File (Microsoft Excel File), 7 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Relevant training parameters of the word embeddings. All the other parameters are set to default values. Two state-of-the-art word embedding models, namely word2vec, and GloVe, have been trained with all the comments and submissions in our subreddits dataset. After a-posteriori validation by a domain expert in terms of topical coherence, we choose word2vec as the reference word embedding model.

[XLSX File (Microsoft Excel File), 6 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Cumulative probability of finding n or fewer terms in a sentence for submissions and comments (left panel). Cumulative probability of having n or fewer sentences in a submission or a comment (right panel). Plots refer to the selected subreddit in 2018.

[PNG File, 128 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Odds Ratios of opioid substances and Secondary Routes of Administration. The central line and the bar mark the OR and the 95% confidence interval respectively, while the size of the circle is proportional to the sample of co-mentions. Measures that are not statistically significant ($P > .01$) are reported in gray. Labels on the right axis report the Odds Ratio and the confidence interval.

[PNG File, 694 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Odds Ratios of opioid substances and Drug Tampering Methods. The central line and the bar mark the OR and the 95% confidence interval respectively, while the size of the circle is proportional to the sample of co-mentions. Measures that are not statistically significant ($P > .01$) are reported in gray. Labels on the right axis report the Odds Ratio and the confidence interval.

[PNG File, 577 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Odds Ratios of Secondary Routes of Administration and Drug Tampering Methods. The central line and the bar mark the OR and the 95% confidence interval respectively, while the size of the circle is proportional to the sample of co-mentions. Measures that are not statistically significant ($P > .01$) are reported in gray. Labels on the right axis report the Odds Ratio and the confidence interval.

[PNG File, 598 KB-Multimedia Appendix 7]

References

[Note: The References section is not shown in the image.]


44. De Choudhury M, De S. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. 2014 Presented at: Eighth International AAAI Conference on Weblogs and Social Media; June 1-4, 2014; Ann Arbor, MI.

45. Enes KB, Valadares Brum PP, Oliveira Cunha T, Murai F, Couto da Silva AP, Lobo Pappa G. Reddit weight loss communities: do they have what it takes for effective health interventions? 2018 Presented at: IEEE/WIC/ACM International Conference on Web Intelligence (WI-IAT ’18); December 3-6, 2018; Santiago, Chile.
73. Butler SF, Cassidy TA, Chilcoat H, Black RA, Landau C, Budman SH, et al. Abuse rates and routes of administration of
69. 2014 - Final Rule: Rescheduling of Hydrocodone Combination Products From Schedule III to Schedule II. Drug Enforcement
68. Allan K, Burridge K. Forbidden words: Taboo and the censoring of language. Cambridge, United Kingdom: Cambridge
62. Seabold S, Perktold J. Statsmodels: Econometric and statistical modeling with python. 2010 Presented at: 9th Python in
56. pycld2. Python Software Foundation. URL: https://pypi.org/project/pycld2/ [accessed 2020-10-12]
55. langdetect. Python Software Foundation. URL: https://pypi.org/project/langdetect/ [accessed 2020-12-14] [WebCite Cache ID https://pypi.org/project/langdetect/]
49. Baumgartner J. r/datasets - I have every publicly available Reddit comment for research. Reddit. URL: https://www.reddit.com/r/datasets/comments/3bxlg7/ [accessed 2020-05-27]


Abbreviations

OR: odds ratio
ROA: routes of administration
UMAP: uniform manifold approximation and projection

©Duilio Balsamo, Paolo Bajardi, Alberto Salomone, Rossano Schifanella. Originally published in the Journal of Medical Internet Research (http://www.jmir.org), 04.01.2021. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on http://www.jmir.org/, as well as this copyright and license information must be included.