Original Paper

Identifying Stigma Phenotypes in Social Media Narratives of Substance Use: Observational Study

Lexie Chenyue Wang[1], MSc; Kenneth C Pike[2], PhD; Mike Conway[3], MSc, PhD; Annie T Chen[4], MSIS, PhD

[1]Department of Linguistics, University of Washington, Seattle, WA, United States

[2]Office of Nursing Research, University of Washington, Seattle, WA, United States

[3]School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

[4]Department of Biomedical Informatics and Medical Education, School of Medicine, University of Washington, Seattle, WA, United States

Corresponding Author:
Annie T Chen, MSIS, PhD
Department of Biomedical Informatics and Medical Education
School of Medicine
University of Washington
Department of Biomedical Informatics & Medical Education, University of Washington
Box 358047
Seattle, WA, 98195
United States
Phone: 1 206 543 2259
Email: atchen@uw.edu

Abstract

Background: Individuals with substance use problems experience stigma in different contexts. Identifying characteristic situations in which stigma occurs or manifests—stigma phenotypes—can serve as important leverage points for future intervention.

Objective: This paper aims to (1) identify stigma phenotypes expressed in social media narratives related to substance use stigma and (2) explore the similarities and differences between the stigma phenotypes from a social ecological perspective.

Methods: We collected Reddit posts pertaining to 3 substances—alcohol, cannabis, and opioids. We performed feature engineering using a combination of content analysis, machine learning, and keyword-based methods to predict variables at different levels of the social ecological framework. Leveraging these features, we applied the fuzzy c-means clustering algorithm on the subset of posts containing stigma to extract stigma phenotypes, where a phenotype is defined by four main dimensions: (1) the stigma mechanism present (eg, internalized stigma, anticipated stigma, or enacted stigma), (2) the substance used (eg, alcohol, cannabis, or opioids), (3) the settings involved (eg, work, school, or home), and (4) the actors involved (eg, family, friends, or partners). Finally, we used Kruskal-Wallis and Dunn post hoc tests to examine the differences between stigma phenotypes with respect to specific ecological factors.

Results: We derived 7 stigma phenotypes from stigma-related posts by 8627 authors. The phenotypes can be categorized into 4 groups: internalized stigma–only, anticipated stigma, enacted stigma–only, and mixed-stigma phenotypes. Narratives on internalized stigma phenotypes focused on the self, with minimal reference to settings and actors. One phenotype focused on anticipated stigma and was characterized by a high proportion of opioid use mentions (707/1217, 58.09% of the authors) and references to the health care setting (647/1217, 53.16% of the authors). Posts associated with the enacted stigma–only phenotypes included substantial representation of settings and actors. Narratives in the mixed-stigma phenotypes often involved more than one stigma mechanism, setting, and actor, with home and family being the most salient factors. The phenotypes differed from one another with respect to social ecological factors, including loneliness and social isolation, use of treatment services, presence of health care providers, community and support groups, society, and legalization.

Conclusions: These findings provide valuable insights that help inform the design and development of interventions targeted at different stigma phenotypic groups from a social ecological perspective.

(J Med Internet Res 2025;27:e68695) doi: 10.2196/68695

KEYWORDS

stigma; substance use; social media; machine learning; social ecological

Introduction

Background

Stigma, a term originating in ancient Greece, was initially used to refer to physical marks on an individual's body that signified their status as a criminal, slave, or traitor [1]. In contemporary times, the concept of stigma can include labeling, stereotyping, social exclusion, status loss, and discrimination against individuals [2]. People with substance use disorders frequently experience stigma, which can adversely impact different aspects of their lives, including physical and mental well-being, employment opportunities, social status, relationships, health care, and more [3,4]. Specifically, the negative impacts may include reduced social functioning [5], social disapproval and labeling [6], distrust in health care providers and reluctance to seek treatment [7,8], delayed recovery [9], and insufficient health care [9].

Understanding factors that shape substance use stigma experiences is necessary for effective intervention design for stigma reduction, substance use recovery, and improvement of overall mental and physical health. In this study, we focused on 3 groups of factors: substances, stigma mechanisms, and ecological factors. We focused on 3 substances—alcohol, cannabis, and opioids. Due to differences in legality, societal acceptance, and other considerations, we expected the nature of stigma-related experiences with each substance to differ.

We studied 3 stigma mechanisms—enacted, anticipated, and internalized stigma, as defined in the stigma framework [10,11]. Enacted stigma refers to direct experiences of stereotyping and discrimination enacted by another individual because of a stigmatized attribute [12]. One example is being called a "drug addict" by another person. Anticipated stigma is the expectation that others might hold stereotypical and discriminatory views toward the self [12], for example, hiding liquor bottles from family members due to fear of judgment. Internalized stigma, also known as self-stigma, is the experience of devaluation, shame, unworthiness, and guilt caused by applying the negative beliefs of society concerning a stigmatized attribute to the self [5,13].

Substance use stigma experiences are largely influenced by an individual's interaction with their immediate physical, social, and cultural environments and the people around them [1,2,14,15]. Thus, it is valuable to adopt an ecological approach to examine the significant determinants at different levels of influence [16]. While similar approaches have been used in research concerning other groups, particularly persons who experience HIV stigma [17,18], there is a notable gap in understanding stigma in the context of substance use from a social ecological perspective.

We adapted the social ecological framework (SEF), which is frequently used for health promotion and intervention design [16], to hypothesize key constructs related to substance use stigma. The SEF is particularly suited for this research because it offers a comprehensive perspective that captures the influence of factors across multiple levels, including individual, interpersonal, contextual, community, and societal. This makes it a robust framework for interpreting health behaviors and developing interventions at the population level [16]. In addition, the SEF is a flexible framework that can be applied to focus on specific contexts, often referred to as settings [19,20]. Adapting from definitions in previous research, we defined *behavior settings* to be physical and conceptual settings where substance use stigma takes place [19,21]. We identified 5 behavior settings at the contextual level of influence in the SEF—home, school, work, leisure, and health care—that are potentially insightful for examining stigma manifestations and 5 *actors*—family, partners, friends, coworkers, and health care providers—who may play a role in these scenarios.

Using the ecological features, we aimed to identify characteristic ways in which stigma manifests, or *stigma phenotypes*, from social media data to better understand experiences of stigma in diverse social ecological contexts. We conceptualized a *stigma phenotype* as a distinct combination of features along 4 key dimensions: stigma mechanism, substance, setting, and actor. We sought to identify the patterns of co-occurrence of *stigma mechanisms* (internalized, anticipated, and/or enacted) with respect to different *substances* (alcohol, cannabis, and/or opioids) within diverse *settings* (home, work, school, leisure, and/or health care) and involving different *actors* (family, partners, and/or friends). The identification of stigma phenotypes can, in turn, provide valuable insights that inform the design of customized interventions tailored to different target groups.

There is a significant gap in the literature regarding the use of extensive, naturally occurring data to analyze experiences of substance use stigma. Most of the extant literature on this topic has used survey or interview methods, typically involving small to moderate sample sizes [9]. Social media enables the collection of extensive, naturally occurring data, which is not easily achievable in laboratory settings [22]. Reddit has been frequently used in substance use research as it provides a rich body of pseudoanonymous, user-generated content and features subreddits, or discussion forums, dedicated to topics related to substance use [23-26]. Given the volume of data, we leveraged natural language processing (NLP) and machine learning (ML) methods to derive the defining features of stigma phenotypes.

Objectives

The purpose of this study was to examine stigma phenotypes in social media data to ensure a comprehensive and nuanced understanding of how substance use stigma operates within and across different ecological contexts. Cluster analysis is an unsupervised ML method that can reveal hidden patterns within a dataset and groups data points into clusters based on identified patterns [27]. Due to its ability to identify natural groupings within datasets [27], cluster analysis methods have been frequently used in biomedical research to extract disease phenotypes [28-30]. We performed cluster analysis to extract stigma phenotypes delineated by stigma mechanisms, substances, and ecological factors (behavior settings and actors) from an extensive social media dataset. In addition, we sought to investigate the differences between stigma phenotypes with respect to more granular ecological factors. We propose the following research questions (RQs):

1. What stigma phenotypes (defined by stigma mechanism, substance, setting, and actors) are reflected in social media data?
2. What additional ecological factors might influence these stigma phenotypes?

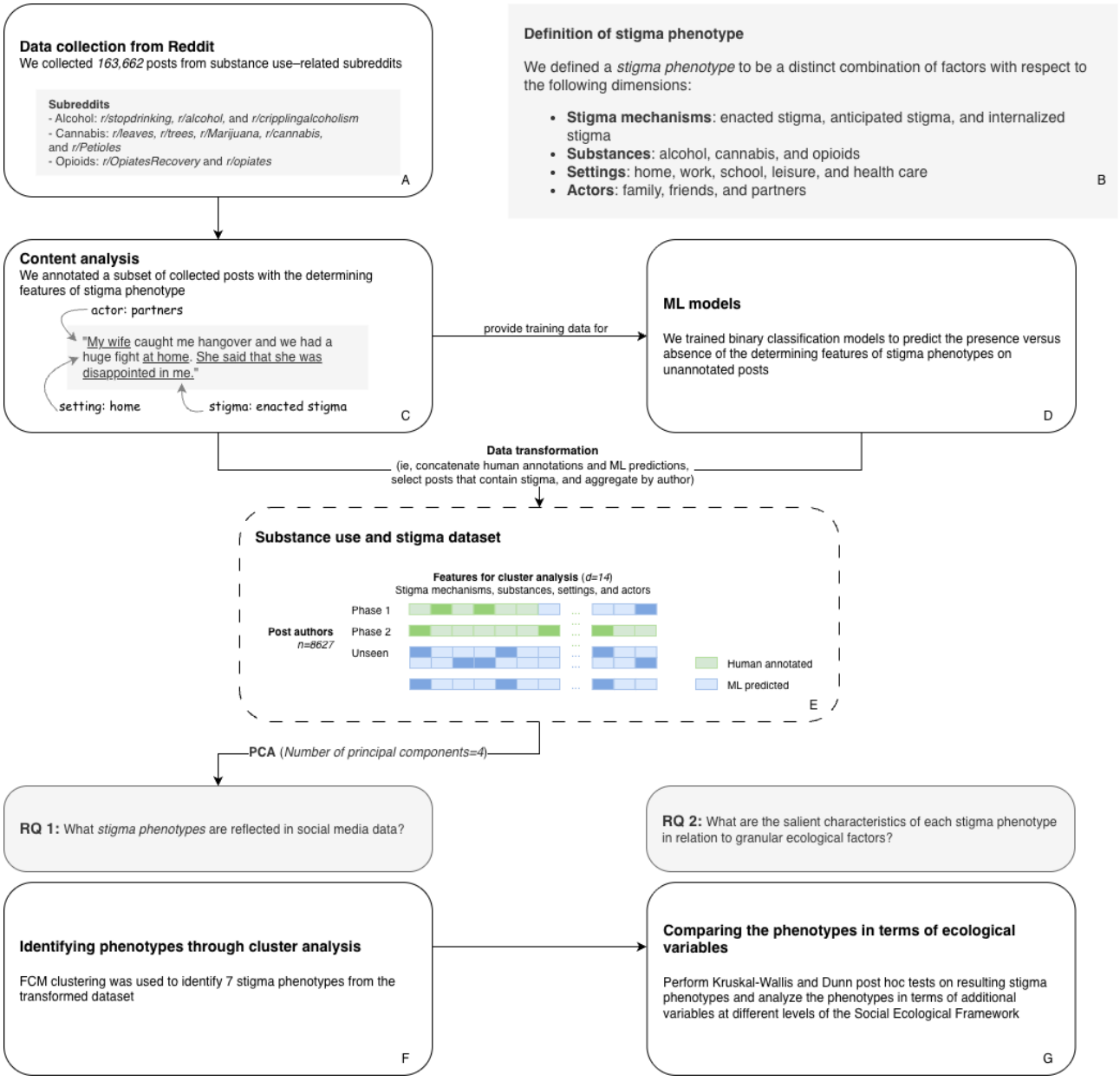
The identification of stigma phenotypes offers valuable insights in guiding the design and development of targeted interventions that address the specific needs and challenges faced by different groups experiencing substance use stigma. By leveraging these phenotypes, targeted interventions can acknowledge the heterogeneity in stigmatizing experiences and deliver support that is contextually relevant to individuals' lived experiences.

Methods

Data Collection

We collected a total of 163,662 Reddit posts published between January 1, 2013, and December 31, 2019, from 10 subreddits (Figure 1, box A) using the Pushshift application programming interface [31]. Each subreddit focuses on 1 of our 3 substances of interest. Some subreddits (eg, r/OpiatesRecovery, r/leaves, and r/stopdrinking) are recovery focused, whereas the others (eg, r/opiates, r/cannabis, and r/cripplingalcoholism) are not. We only collected thread-initiating posts as they tend to contain richer information than their replies [32]. More information about the data collection process is discussed in our prior work [33].

Figure 1. Study flowchart. FCM: fuzzy c-means; ML: machine learning; PCA: principal component analysis; RQ: research question.



Feature Selection

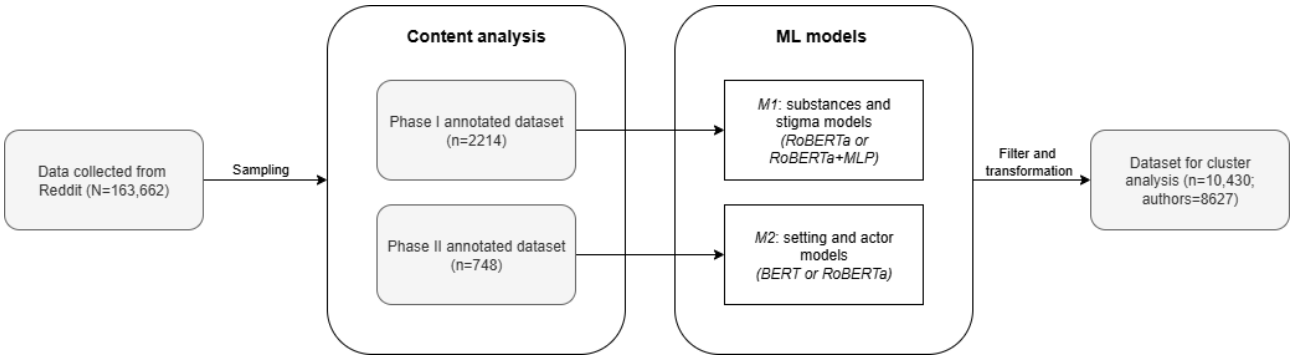
A total of 14 features (ie, factors) were included in the cluster analysis (Figure 1, box B). There were 4 groups of features: substances (alcohol, cannabis, and/or opioids), stigma mechanisms (internalized stigma, anticipated stigma, and/or enacted stigma), settings (home, school, work, health care,

and/or leisure), and actors (family, partners, and/or friends). The settings refer to both physical and conceptual settings; for example, *home* refers to both the presence or absence of the physical space and the idea of home. As such, if an individual mentioned housing insecurity in the passage (eg, being forced to leave their home, which could be important in the context of substance use [34]), this incident would be captured in the data. Specifically, *leisure* was defined as recreational situations in which a social element is present or implied. Each feature could take 1 of 2 values: 0 (absent) or 1 (present).

Feature Engineering

We used NLP methods to develop ML classifiers [35,36] that predicted the presence of each variable. We then used the classifiers to generate predictions on unseen data to be used for subsequent analyses. The feature engineering pipeline consisted of 2 main parts: content analysis (Figure 1, box C) and model development (Figure 1, box D). We conducted 2 phases of content analysis that corresponded to 2 sets of ML models, each serving as a foundation to derive different features necessary for the subsequent cluster analysis (Figure 2).

Figure 2. Feature engineering flowchart. BERT: bidirectional encoder representations from transformers; MLP: multilayer perceptron.



The phase 1 annotated dataset (n=2214 posts) was used to train models that predict the presence versus absence of the substances and stigma mechanisms [33], referred to as M1 models. The posts in this dataset were sampled using the keyword sampling approach described by Chen et al [12] to select posts that likely contained stigma. Posts were annotated with substances and stigma mechanisms. Further details on the operationalization of the stigma mechanisms are provided in Multimedia Appendix 1.

All the M1 models involved fine-tuning a pretrained RoBERTa model [35]. As stigma-related information is often not explicitly stated in text and requires interpretation beyond surface-level content, we incorporated an additional multilayer perceptron into the M1 models. The multilayer perceptron leveraged a variety of features, including term frequency–inverse document frequency weighted n-grams, the National Research Council Canada Emotion Intensity Lexicon [37], the WordNet-Affect Lexicon [38], Linguistic Inquiry and Word Count features [39], and other handcrafted features [33]. We report the F_1-score as the primary evaluation metric for all models because it balances precision and recall, which is appropriate given that we do not have a strong preference between the 2 metrics. M1 substance models achieved an average F_1-score of 0.96 (SD 0.01), whereas M1 stigma models achieved an average F_1-score of 0.80 (SD 0.06). Further details on M1 model design, development, and evaluation can be found in our prior work [33].

In phase 2, we sought to identify a wider range of phenomena, including ecological features such as settings and actors. The phase 2 annotated dataset (n=748 posts) included 496 posts from the phase 1 annotated dataset and 252 posts added through quota sampling involving a balanced representation of each substance and stigma type to enrich the representation of stigma within the dataset. This dataset was used to develop classifiers that predicted settings and actors, referred to as M2 models. M2 models used only transformer-based architectures—either bidirectional encoder representations from transformers (BERT) or RoBERTa. To address challenges such as limited training data size and imbalanced data, we implemented the BERTprepend [40] data augmentation method on the training set. We then fine-tuned pretrained BERT and RoBERTa models [41,42] on the augmented dataset and selected the better-performing model for each feature. BERT achieved better performance for home, leisure, work, health care, and friends, and RoBERTa was better at predicting school, family, and partners. The M2 BERT and RoBERTa models achieved an average F_1-score of 0.83 (SD 0.03) and 0.81 (SD 0.16), respectively, across the held-out folds of 5-fold cross-validation. Table 1 presents a summary of the model architectures, predicted features, and model performances.

Table 1. Overview of machine learning–based binary classification models used in feature engineering, showing model architectures, target variables, training and evaluation datasets, and model performance measured using the average F1-score.

Model and architecture	Target variables	Dataset	F_1-score, mean (SD)
M1[a] models			
RoBERTa+MLP[b]	Stigma mechanisms: enacted stigma, anticipated stigma, and internalized stigma	Phase 1 annotated dataset	0.80 (0.06)
RoBERTa	Substances: alcohol, cannabis, and opioids	Phase 1 annotated dataset	0.96 (0.01)
M2[c] models			
BERT[d]	Settings: home, leisure, work, and health care; actors: friends	Phase 2 annotated dataset	0.83 (0.03)
RoBERTa	Settings: school; actors: family and partners	Phase 2 annotated dataset	0.81 (0.16)

[a]M1: ML models that predict stigma mechanisms and substances.

[b]MLP: multilayer perceptron.

[c]M2: ML models that predict settings and actors.

[d]BERT: bidirectional encoder representations from transformers.

Identifying Phenotypes Through Cluster Analysis

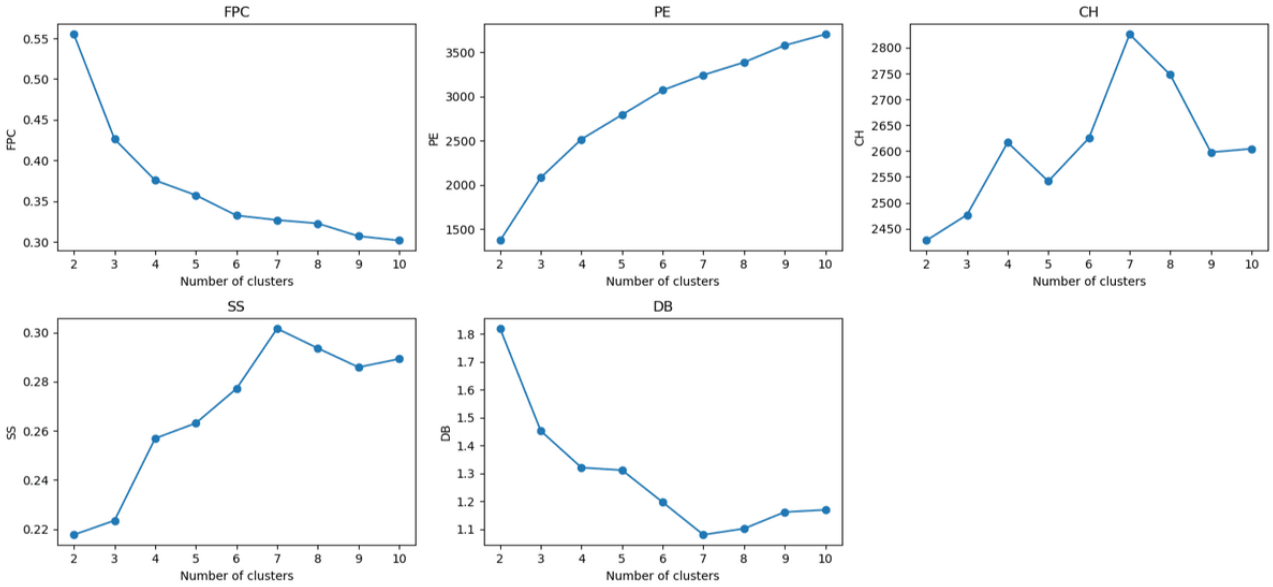
As most collected posts did not contain stigma, we filtered the cluster analysis dataset to retain only stigma-positive posts, ensuring that the analyses were exclusively focused on stigma-related content. Once the features were derived, we restructured the dataset from post level to author level, aggregating the values of each feature across all posts written by the same person (Figure 1, box E). This procedure was performed to focus on understanding the stigma-related experiences of each individual on a holistic level.

To address RQ 1 (Figure 1, box F), we used the fuzzy c-means (FCM) clustering algorithm to derive stigma phenotypes from the data [43]. As we explored the appropriate clustering techniques, we evaluated a variety of methods, including but not limited to centroid-based, hierarchical clustering and fuzzy or soft clustering methods [44]. Unlike hard clustering methods that assign each case to a single cluster, FCM offers flexibility in handling ambiguous and complex data through partial cluster membership [45], making it particularly well suited for the analysis of substance use stigma, which often manifests in layered and nuanced ways.

With 14 features, we encountered the "curse of dimensionality" [46], meaning that an inverse relationship between meaningful differentiation between clusters and the number of features (or dimensions) was observed. To address this problem, we used principal component analysis (PCA) to condense the feature space down to 4 dimensions while preserving >50% of the cumulative explained variance in our data.

We used a range of validity indexes to determine the optimal number of clusters (c^*) [47]—fuzzy partition coefficient, partition entropy, Calinski-Harabasz index, silhouette score, and Davies-Bouldin index (Figure 3). The fuzzy partition coefficient and partition entropy favored the smallest value (c^*=2), which was not ideal for understanding complex and fuzzy stigma phenotypic patterns. In contrast, the Calinski-Harabasz index, silhouette score, and Davies-Bouldin index metrics, which measure clustering quality based on intercluster separation and intracluster compactness [48], were all in favor of c^*=7 compared to other numbers. In summary, we used the FCM clustering algorithm to derive 7 stigma phenotypes from a PCA-transformed dataset with 4 principal components reduced from 14 raw features.

Figure 3. Validity indices for determining the optimal number of clusters c*. Five validity indices (fuzzy partition coefficient [FPC], partition entropy [PE], Calinski-Harabasz index [CH], silhouette score [SS], and Davies-Bouldin index [DB]) evaluated across 2 to 10 clusters to identify the optimal c* for deriving stigma phenotypes.

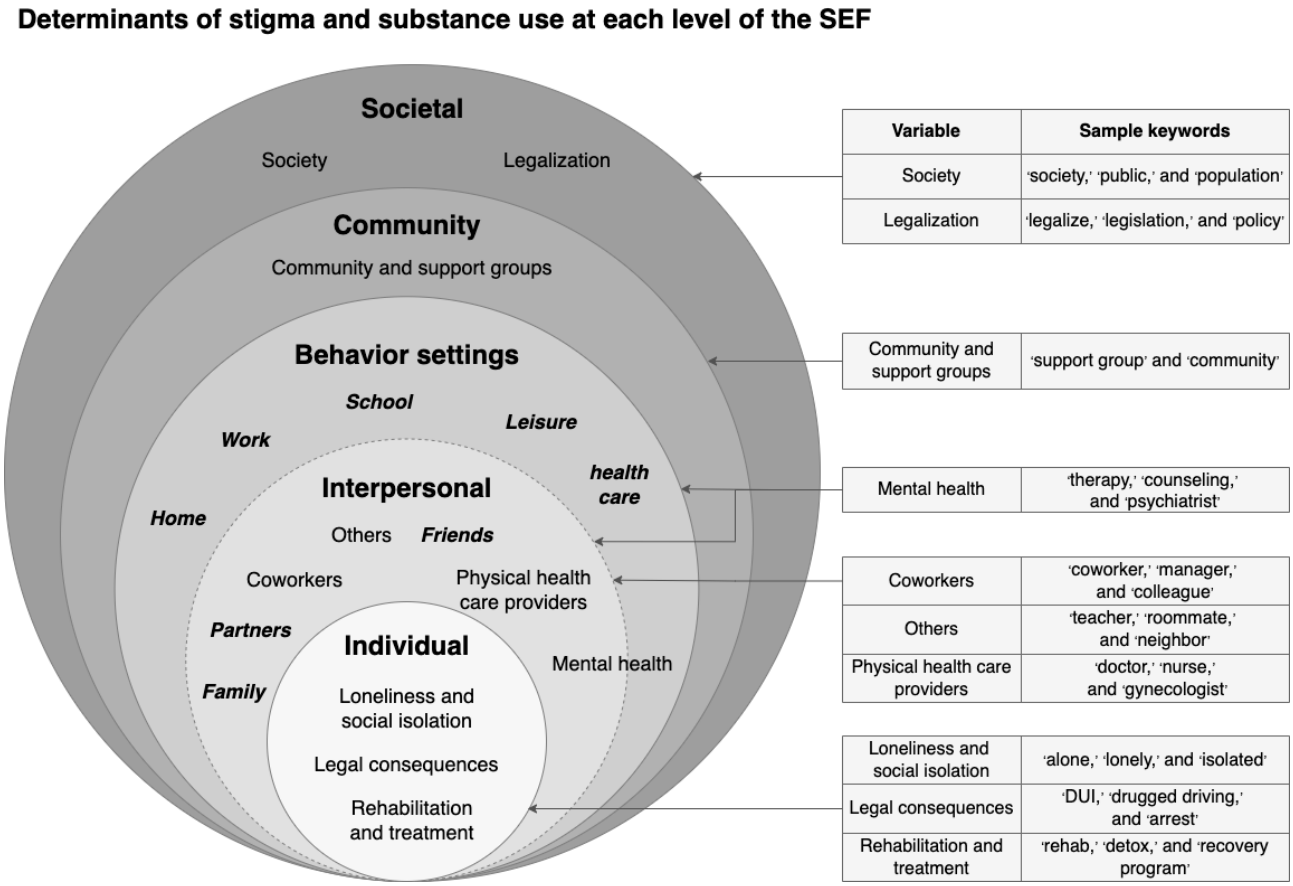


Comparing the Phenotypes in Terms of Ecological Variables

In cluster analysis, it is common to include the dimensions that the researcher considers to be the defining characteristics, or the primary features of interest, as the variables to be clustered upon. Post hoc analyses are performed subsequently to better understand factors that may influence the defining characteristics [49].

In this study, the primary characteristics of interest were the settings and actors that might be involved in stigma-related experiences, and thus, these served as the focus of RQ 1. However, also consistent with an ecological view, human activity is influenced by a myriad of other factors that intersect at different levels. Thus, for RQ 2 (Figure 1, box G), we sought to identify additional factors that shape substance use stigma experiences at different levels of influence guided by the SEF [16,50] and compare their relative prominence with respect to the phenotypes (Figure 4).

Figure 4. Social ecological framework (SEF) for determinants of substance use stigma at different levels of influence and sample keywords used for keyword-based statistical analysis.



We used a keyword-based approach to systematically capture ecological variables. The full list of keywords can be found in Multimedia Appendix 2. We developed keyword lists and then normalized keyword counts by calculating the percentage of sentences containing the keywords, which allowed for fair and meaningful comparisons across posts of varying lengths.

At the individual level, *loneliness and social isolation* were related to substance use and stigma [51-53]. Through content analysis, a systematic method of assigning codes to text segments [54], we observed that *legal consequences* and *rehabilitation and treatment* were frequent themes in the dataset. For example, individuals can be stigmatized for driving under the influence of a substance, and people may choose to engage in rehabilitation, treatment, or therapy; capturing these could facilitate a more comprehensive view of the challenges that those experiencing substance use stigma might face.

At the interpersonal level, the actors included *family*, *partners*, and *friends*, which were prominent actors in the dataset and were included in cluster analysis. Interactions with other actors such as *coworkers*, *health care providers*, and *others* (eg, servers, roommates, and barbers) can also play critical a role in the experiences of a stigmatized individual and, thus, were included in the post hoc analyses.

Behavior settings are physical and conceptual settings in which stigma and substance use experiences take place [19]. *Home*, *work*, *school*, *leisure*, and *health care* are prominent settings that were included in the cluster analysis. It is important to note that the interpersonal variables and behavior settings are often interconnected but do not completely overlap. For example, one can experience stigma at home, but the stigma is not necessarily enacted by family or partners. *Mental health care* sits at the intersection of the interpersonal level and behavior settings as it covers both the mental health care setting and mental health care providers.

At the community level, we included the variable *community and support groups*, which covers both online and offline communities and support groups such as Alcoholics Anonymous. This is an important variable as communities (primarily online communities, referring to subreddits in this dataset) and support groups can help individuals succeed at substance use recovery [55,56]. At the societal level, *legalization* of substances is an important factor to consider as individuals who use illicit substances often experience higher levels of stigma [57]; *society* itself is also an actor.

We used statistical tests to examine the differences among the 7 stigma phenotypes regarding these granular variables. As interim multinomial logistic regression results showed minimal to no interaction effects, we independently analyzed differences between the clusters on the granular ecological variables using the nonparametric Kruskal-Wallis test. To account for multiple comparisons, we performed the Dunn post hoc test with Bonferroni adjustment. These methods are ideal for comparing the phenotypes on nonnormally distributed keyword-based variables [58]. The full set of test statistics can be found in Multimedia Appendix 3.

Ethical Considerations

In this paper, we use synthetic or paraphrased quotes derived from original data to provide examples of key themes in the data. We use synthetic quotes to protect author privacy as direct quotes from online sources could be traced back to their original posts [59]. As we developed the synthetic quotes, we worked to ensure that the quotes were representative of the raw data and accurately reflected the language, meaning, and connotation of the original content [60]. This study received ethics approval from the University of Washington Human Subjects Division (STUDY00015737) and the University of Melbourne Human Research Ethics Committee (2023-25512-48127).

Results

Sample Statistics

Our sample comprised 10,430 stigma-related posts representing 8627 authors. Table 2 shows the descriptive statistics of the dataset at both the author and post levels. The numbers reported in the text pertain to the author level.

Table 2. Sample characteristics regarding feature presence at the author and post levels based on predictions from the machine learning models.

Feature	Author level (n=8627), n (%)	Post level (n=10,430), n (%)
Substances		
Alcohol	4571 (52.98)	5510 (52.83)
Cannabis	3156 (36.58)	3680 (35.28)
Opioids	1288 (14.93)	1552 (14.88)
Stigma mechanisms		
Enacted stigma	3390 (39.3)	3832 (36.74)
Anticipated stigma	2767 (32.07)	3031 (29.06)
Internalized stigma	4823 (55.91)	5560 (53.31)
Settings		
Home	3590 (41.61)	4003 (38.38)
Work	2753 (31.91)	3038 (29.13)
School	1358 (15.74)	1437 (13.78)
Leisure	2784 (32.27)	3035 (29.1)
Health care	2016 (23.37)	2202 (21.11)
Actors		
Family	3616 (41.91)	4034 (38.68)
Partners	2720 (31.53)	2991 (28.68)
Friends	3013 (34.93)	3244 (31.1)

Among the sample, the proportion of authors mentioning each substance varied, with 52.98% (4571/8627) mentioning alcohol, 36.58% (3156/8627) mentioning cannabis, and 14.93% (1288/8627) mentioning opioids. Most authors (7690/8627, 89.1%) indicated use of a single substance within the scope of the posts captured in this dataset.

Internalized stigma (4823/8627, 55.91%) was the most prominent stigma mechanism, more common than enacted stigma (3390/8627, 39.3%) and anticipated stigma (2767/8627, 32.07%). Among the ecological variables, home (3590/8627, 41.61%) and family (3616/8627, 41.91%) were more prominent, whereas school (1358/8627, 15.74%) and health care (2016/8627, 23.37%) were less common.

The Stigma Phenotypes

Overview

We identified 7 distinct stigma phenotypes from this dataset. Table 3 summarizes the average post length of each phenotype.

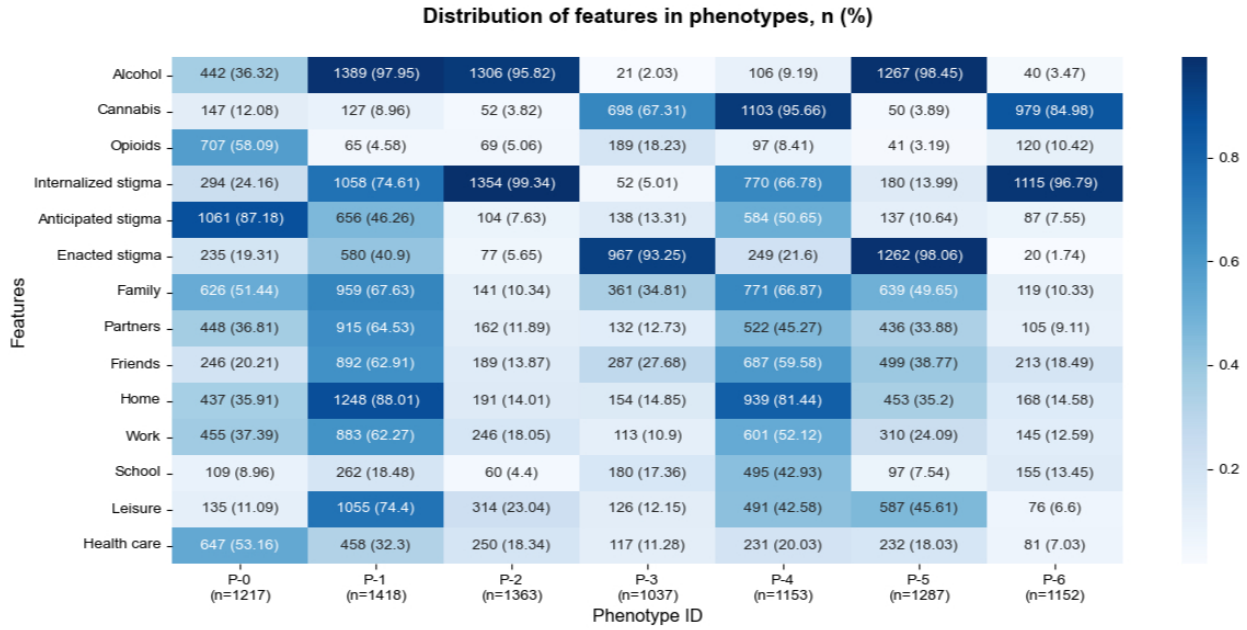
Table 3. Post length statistics by stigma phenotype.

Phenotype ID	Authors (n=8627), n (%)	Number of words, mean (SD)	Number of words, median (IQR)
0	1217 (14.1)	309 (322)	223 (235)
1	1418 (16.44)	381 (277)	310 (246)
2	1363 (15.8)	177 (170)	137 (131)
3	1037 (12.02)	233 (328)	152 (223)
4	1153 (13.37)	433 (351)	344 (307)
5	1287 (14.92)	259 (292)	199 (182)
6	1152 (13.35)	200 (185)	154 (164)

Figure 5 illustrates the distribution of features in each stigma phenotype, with darker cells indicating higher presence and lighter cells representing lower presence. Even though PCA was used to reduce dimensionality, the interpretability of transformed components is conveyed through their alignment with original dimensions. Thus, the reporting of statistics in the characterization of stigma phenotypes was guided by the input features that were used to derive the principal components. The stigma mechanisms and substances were the primary defining characteristics of the clusters, and the setting and actor features contributed to the differentiation between clusters in a more nuanced manner. The phenotypes were as follows:

1. *opioids/alcohol–anticipated stigma–health care* (P-0)
2. *alcohol–mixed stigma–more settings and actors* (P-1)
3. *alcohol–internalized stigma–few settings and actors* (P-2)
4. *cannabis/opioids–enacted stigma–few settings and actors* (P-3)
5. *cannabis–mixed stigma–more settings and actors* (P-4)
6. *alcohol–enacted stigma–few settings and actors* (P-5)
7. *cannabis–internalized stigma–few settings and actors*

Figure 5. Distribution of substances, stigma mechanisms, actors, and settings across 7 identified stigma phenotypes. The heatmap is colored based on the percentage of feature presence in each phenotype, with darker colors indicating stronger presence.



The 7 phenotypes can be broadly categorized into 2 groups distinguished by shared presence or absence of certain features: those characterized by a single prominent stigma mechanism (P-0, P-2, P-3, P-5, and P-6) and those displaying moderate to high prominence of multiple stigma mechanisms (P-1 and P-4). Among the single-stigma group, there was a distinct separation of the 3 stigma types, with P-2 and P-6 characterized by internalized stigma, P-0 characterized by anticipated stigma, and P-3 and P-5 characterized by enacted stigma.

Single-Stigma Phenotypes

Phenotype 2 (P-2; 1363/8627, 15.8% of the authors) and phenotype 6 (P-6; 1152/8627, 13.35% of the authors) were characterized by a high presence of internalized stigma and low presence of anticipated and enacted stigma. The main difference between P-2 and P-6 was the predominant substance—P-2 authors primarily used alcohol (1306/1363, 95.82%), whereas P-6 authors mainly used cannabis (979/1152, 84.98%), with a small proportion using opioids (120/1152, 10.42%). The posts were relatively short (Table 3), and settings and actors were rarely mentioned in both phenotypes. The narratives in P-2 and P-6 were usually centered on the posters' internal feelings and thoughts, without providing much detail about the context:

> *I am embarrassed by my drunk actions. I feel like I'm not capable of anything.* [P-2; internalized stigma; alcohol]

There were 2 phenotypes with enacted stigma as the distinguishing characteristic (P-3, 1037/8627, 12.02% of the authors and P-5, 1287/8627, 14.92% of the authors). P-5 was characterized by alcohol use (1267/1287, 98.45%), whereas P-3 was mainly associated with cannabis use (698/1037, 67.31%), with some authors (189/1037, 18.23%) reporting opioid use. P-3 and P-5 also differed in terms of the involvement of ecological features. Settings and actors such as family, friends, partners, home, and leisure were common themes in P-5; however, the presence of settings and actors was lower in P-3, except for school. In both enacted stigma phenotypes, it was common for the narratives to center on specific situations, with family and friends being frequently mentioned:

> *My wife caught me hung over and we had a huge fight at home. She said that she was disappointed in me, and it really hurts.* [P-5; enacted stigma; alcohol]

Phenotype 0 (P-0; 1217/8627, 14.11% of the authors) was the only phenotype in which anticipated stigma was the leading stigma mechanism (1051/1217, 86.36%). Internalized stigma and enacted stigma were also experienced by some individuals but to a lesser extent. The substances used by P-0 authors varied, with 58.09% (707/1217) reporting opioid use, 36.32% (442/1217) reporting alcohol use, and 12.08% (147/1217) reporting cannabis use. Most settings and actors were moderately common, except for school and leisure.

Notably, P-0 was the only phenotype in which health care was the most prominent setting (647/1217, 53.16%). In many cases, anticipated stigma and health care coexisted without being directly related; however, there were scenarios in which the connections between the 2 features were evident. Some individuals felt the need to seek help from therapists because they hid their substance use problems from people around them, and some others anticipated stigma from health care providers:

> *I need to talk to a therapist about my situation because there's no one else I can talk to. My wife and my family can't find out that I'm using again.* [P-0; anticipated stigma; health care]

> *How do I get my Dr to prescribe more pain killers without me appearing like a drug seeker?* [P-0; anticipated stigma; health care]

Mixed-Stigma Phenotypes

There were 2 phenotypes in which authors experienced multiple stigma mechanisms. Phenotype 1 (P-1; 1418/8627, 16.44% of the authors) was characterized by alcohol use (1389/1418, 97.95%). In P-1, a total of 74.61% (1058/1418) of the authors mentioned internalized stigma, 46.26% (656/1418) mentioned anticipated stigma, and 40.9% (580/1418) mentioned enacted stigma. Phenotype 4 (P-4; 1153/8627, 13.37%) was characterized by cannabis use (1103/1153, 95.66%). In total, 66.78% (770/1153) of P-4 authors discussed internalized stigma, 50.65% (584/1153) mentioned anticipated stigma, and 21.6% (249/1153) discussed enacted stigma. Narratives in P-1 and P-4 were lengthier (P-1: mean 381, SD 277 words; P-4: mean 433, SD 351 words) and more likely to involve settings and actors. The interconnection between different stigma mechanisms was evident in P-1 and P-4 posts. For example, many authors discussed feelings of internalized stigma and past experiences of enacted stigma:

> *How do you all deal with the constant feeling of shame? I keep replaying what happened the other day: I got drunk, blacked out, and ended up vomiting in my friend's car. I could tell my friend was judging me, and I totally deserve it. I feel so guilty and ashamed, and I can't stop thinking about it. I really want to quit drinking.* [P-1; alcohol; enacted stigma; internalized stigma; friends]

In addition, there were notable differences between phenotypes characterized by different substances. In the internalized stigma–only, enacted stigma–only, and mixed-stigma groups, the school setting was consistently more prominent in the cannabis phenotypes (P-6, P-3, and P-4) compared to the alcohol phenotypes (P-2, P-5, and P-1). Although not many authors explicitly discussed incidents occurring at school, many mentioned that they began smoking cannabis while they were students. Conversely, the leisure setting was consistently more prominent in the alcohol phenotypes than in their cannabis and opioid counterparts. Authors frequently referred to leisure settings to provide context for their stories, but there were also instances in which they discussed stigma-related challenges encountered in leisure contexts:

> *I'm excited to attend my friend's wedding, but I'm worried about what I should say when they offer me drinks. None of them knows that I am an alcoholic and have been trying to stay sober.* [P-1; alcohol; anticipated stigma; leisure]

Comparing the Phenotypes in Terms of Ecological Variables

We conducted Kruskal-Wallis and Dunn post hoc tests with Bonferroni correction to compare and analyze the phenotypes with respect to granular ecological variables. In this section, we focus on the most meaningful comparisons, which are shown in Table 4. The unadjusted and adjusted *P* values for the comparisons between each combination of phenotypes can be found in Multimedia Appendix 3.

Table 4. Notable statistically significant pairwise comparisons from Kruskal-Wallis and Dunn post hoc tests with Bonferroni correction (adjusted P<.05)[a].

Level and variable	Notable comparisons
Individual	
Loneliness and social isolation	• P-1 and P-4>other phenotypes[b]
Rehabilitation and treatment	• P-0>other phenotypes • P-1>other phenotypes except for P-0 • P-5>P-3, P-4, and P-6
Legal consequences	• P-3>other phenotypes
Interpersonal	
Coworkers	• P-1<other phenotypes
Physical health care providers	• P-0>other phenotypes
Others	• P-1>P-0, P-2, P-3, P-5, and P-6 • P-4>P-0, P-2, P-3, and P-6 • P-5>P-0, P-2, and P-6
Behavior settings	
Mental health care	• P-0, P-1, and P-4>other phenotypes
Community	
Community and support groups	• P-1>other phenotypes
Society	
Society	• P-3>P-0, P-2, P-5, and P-6 • P-1>P-2, P-5, and P-6
Legalization	• P-3>other phenotypes

[a]P-0 to P-6 denote the stigma phenotypes derived from cluster analysis. See text above for pattern summaries.
bOther phenotypes refers to the set of phenotypes not explicitly listed on the left side of the comparison. For example, in P-1 and P-4>other phenotypes, other phenotypes refers to P-0, P-2, P-3, P-5, and P-6.

In Table 4, we observe that, at the individual level, the mixed-stigma phenotypes (P-1 and P-4) exhibited significantly higher levels of *loneliness and social isolation* than other phenotypes (Table S1 in Multimedia Appendix 3). The notation *P-1 and P-4>other phenotypes* indicates that phenotype 1 was significantly greater with respect to this variable when individually compared to the other phenotypes; it does not imply that P-1 and P-4 are significantly different from each other. *Rehabilitation and treatment* was highest in P-0—characterized by opioid and alcohol use, anticipated stigma, and health care—followed by 2 other alcohol phenotypes, P-1 and P-5 (Table S2 in Multimedia Appendix 3). *Legal consequences* were notably higher in P-3 (Table S3 in Multimedia Appendix 3), which is characterized by cannabis and opioid use, enacted stigma, and relatively low presence of the settings and actors studied:

> *I got pulled over a couple days ago. The cops thought I was high, but I was completely sober. Then they searched me without even asking for my permission.* [P-3; legal consequences]

At the interpersonal level, *coworkers* was significantly higher in P-1 (alcohol; mixed stigma) than in other phenotypes (Table S4 in Multimedia Appendix 3). *Physical health care providers* were mentioned more in P-0, where health care was the most prominent setting (Table S5 in Multimedia Appendix 3). *Others* were most frequently mentioned in P-1, followed by P-4 (cannabis; mixed stigma) and P-5 (alcohol; enacted stigma; Table S6 in Multimedia Appendix 3). P-1, P-4, and P-5 were phenotypes with higher presence of actors in general. The behavior setting *mental health care* was more prominent in P-0, P-1, and P-4 (Table S7 in Multimedia Appendix 3).

At the community level, P-1 authors mentioned *community and support groups* more frequently than authors in other phenotypes (Table S8 in Multimedia Appendix 3). There was a strong sense of (online) community among P-1 authors:

> *34 days sober. I could really use some words of encouragement to support me through the process, and I will stay active in this thread for accountability. Thanks everyone.* [P-1; alcohol; community and support groups]

Finally, at the societal level, *society* was mentioned more by P-3 and P-1 authors compared to authors in some but not all other phenotypes (P-2, P-5, and P-6; Table S9 in Multimedia Appendix 3). P-3 authors, characterized by opioid or cannabis use, discussed *legalization* much more than those in other groups (Table S10 in Multimedia Appendix 3).

Discussion

Principal Findings

Using the FCM clustering algorithm, we derived 7 stigma phenotypes based on posts containing stigma related to 3 substances. Stigma mechanisms and substances were the key determinants of the phenotypes, and settings and actors influenced the differentiation between clusters in a more subtle way. The 7 stigma phenotypes can be categorized into 4 groups: internalized stigma–only group (P-2 and P-6), anticipated stigma group (P-0), enacted stigma–only group (P-3 and P-5), and mixed-stigma group (P-1 and P-4). This distinction highlights significant differences in the manifestations of the stigma mechanisms individually and in concert.

The narratives in the internalized stigma–only phenotypes were characterized by a focus on internal feelings, with few references to contextual factors. In contrast, the narratives in the enacted and anticipated stigma phenotypes usually involved more settings and actors, which exemplifies the inherently interpersonal nature of these 2 stigma mechanisms [11].

We observed a complex connection among anticipated stigma, health care, and opioid use in the anticipated stigma phenotype (P-0), which was the only phenotype where anticipated stigma, opioid use, and health care were significantly more common than other stigma types, substances, and settings, respectively. Health care providers often exhibit negative attitudes toward patients with substance use disorders [61]; consequently, individuals with substance use problems frequently experience anticipated stigma in health care settings, reducing their willingness to seek help from health care professionals [62,63]. This pattern is particularly pronounced among opioid users, exemplified by behaviors such as attempting to hide their substance use history and avoiding health care interactions [4]. Moreover, the high prevalence of the mental health care setting in this phenotype is likely related to anticipated stigma—individuals often seek or consider seeking help from mental health care providers because they find it difficult to discuss their issues with people around them.

In the mixed-stigma phenotypes (P-1 and P-4), the connections between stigma mechanisms were evident. We observed that past experiences of enacted stigma can lead to or reinforce anticipated and internalized stigma [64]. This is in line with the progressive model of self-stigma [65], which posits 4 stages that eventually lead to diminished self-esteem. This first stage is being aware of the stereotypes, which can be caused by incidents of enacted stigma and may result in anticipated stigma. Individuals may then agree with the stereotypes and apply the stereotypes to themselves, which aligns with internalized stigma. Finally, the internalized stigma that individuals experience may lead to lower self-esteem.

In addition, significantly higher levels of loneliness and social isolation were observed in the mixed-stigma phenotypes. While the underlying mechanisms behind this phenomenon remain unclear, several possibilities can be suggested. In some cases, individuals may isolate themselves voluntarily to avoid social interactions and social stigma [66]. Moreover, loneliness and social isolation may also reinforce both anticipated and internalized stigma as individuals continually worry about how others perceive and react to them [66]. There was also a high presence of the mental health care setting in the mixed-stigma group. Given these results, interventions targeting individuals who experience multiple forms of stigma should address the individuals' feelings of loneliness and social isolation and needs for therapy.

Moreover, substantial differences were evident among the substances studied. The leisure setting was more prominent in alcohol phenotypes than in their cannabis and opioid counterparts. Individuals recovering from alcohol use may feel social pressure to drink and experience stigma in leisure contexts as they attempt to hide their substance-related problems. Therefore, interventions for alcohol users could focus on stigma reduction in social and leisure contexts. The school setting was more prominent in cannabis phenotypes. Previous research has demonstrated that cannabis use is notably prevalent among middle and high school students [67,68], and exposure to cannabis use at school increases the likelihood of students' own cannabis use, but the same pattern does not hold for alcohol [69]. Our finding further solidifies the necessity for targeted interventions designed for school settings [70].

Furthermore, there were distinctive ways in which specific phenotypes stood out. P-0, characterized by opioid and alcohol use, was the only cluster in which health care was the most prominent setting. It was common for P-0 authors to mention chronic pain conditions and the use of opioids for pain management [71]. In addition, there were also cases in which the authors engaged with physical health care providers due to consequences of opioid or alcohol use, such as overdose, severe hangovers, or driving under the influence accidents. These results suggest that interventions in the health care scope should address not only the consequences but also the underlying circumstances of substance use.

In comparison to all other phenotypes, the role of community and support groups was highly emphasized in P-1, which was characterized by alcohol and mixed-stigma mechanisms. P-1 individuals had a strong sense of (online) community—they sought social support from each other and provided support when they could. They also used the online community to hold themselves accountable. In addition, offline support groups such as Alcoholics Anonymous were often mentioned in P-1 narratives. Congruent with our own prior work, individuals expressed concerns and uncertainties about attending offline support groups, shared both positive and negative experiences and insights, encouraged others to join, or discussed why support groups did not work well for them [12]. These findings suggest that online communities such as Reddit can serve as valuable resources for individuals with substance use stigma problems. Offline support groups can also provide support and offer a structured environment for recovery, but they may not be

suitable for everyone. While P-3 and P-5 were both characterized by enacted stigma, the involvement of actors and settings was notably lower in P-3 (cannabis and opioids) than in P-5 (alcohol). This discrepancy might be largely attributed to the prevalence of legal consequences, legalization, and society in P-3. Some individuals may express concerns regarding the legal consequences of their substance use behaviors. In other cases, legal consequences can manifest as a form of enacted stigma. For example, individuals may feel that they were unfairly stopped or searched due to their substance use history or appearing to be under the influence. At the societal level, the legalization of a substance can influence the public's attitudes toward the normalization of its use [6]. Individuals in P-3 actively engaged in discussions about cannabis legalization, covering a range of topics. They expressed their opinions on government policies, shared troubles they faced due to the legal status of the substance they used, and discussed how drug laws disproportionately affect ethnic minority groups. These findings highlight the need for interventions to consider the impacts of legal consequences, substance legalization, society, and the stigma associated with substance use.

Strengths and Limitations

This research has the following contributions and strengths. This study examined the manifestations of substance use–related stigma in a large-scale dataset. Unlike many studies that focus on a specific subgroup of individuals who either use a particular substance or experience a specific type of stigma [72,73], this research adopted a holistic approach to identify stigma phenotypes in a large dataset by including users of different substances and their experiences with different stigma mechanisms. Furthermore, even though NLP and ML techniques [74] and the SEF model [75,76] have been used separately in previous studies to explore health care–related topics, this research introduced a novel methodology that combined all 3. This combined approach ensured the scalability of the research by facilitating the understanding and analysis of a volume of data at a level of granularity that is difficult to achieve with qualitative and survey-based methods.

The limitations of this study are related to the nature of the dataset and sampling bias. The analysis was limited to Reddit data from 10 subreddits, which may not represent broader populations who experience substance use stigma, particularly those who are less engaged in social media. The percentage of authors who mentioned opioid use was significantly lower than the percentage of authors who mentioned the other 2 substances of interest, which could introduce bias in the clustering results. In addition, the results of this study largely depend on what the posters chose to disclose and how much detail they chose to include. In general, we found that phenotypes with longer average post lengths (P-0, P-1, and P-4) tended to have a higher presence of stigma, behavior settings, and actors. In phenotypes with shorter posts (P-2 and P-6), these factors were mostly absent. The relationship among post length, detail, and presence of features was subtle and complex. On the one hand, individuals who wrote longer, more detailed posts may have encountered problems in multiple aspects of their lives; thus, their discussions revolved around multiple settings, actors, and stigma mechanisms. On the other hand, they might simply feel a stronger urge of self-disclosure or have a detail-oriented writing style.

Conclusions

In this study, we introduced a novel combination of NLP, ML, and statistical techniques to examine narratives containing stigma related to substance use from a social ecological perspective. We identified 7 phenotypes from large-scale social media data about substance use and stigma. These phenotypes revealed distinct patterns in terms of stigma mechanisms, substances used, settings and actors involved, and the role of ecological factors across different levels of our framework. The findings of this study provide a framework for understanding stigma as a multidimensional and context-dependent phenomenon. The valuable insights from this research could inform the design and development of interventions targeting different stigma contexts.

Acknowledgments

The research reported in this publication was supported by the National Institute on Drug Abuse of the National Institutes of Health under award R21DA056684. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Special thanks are extended to David Roesler for his development of the M1 machine learning models, to Sharon H Wong for overseeing the content analysis process, and to Shana Johnny and Rahul K Chaliparambil for their efforts in annotation of the data used in this study.

Data Availability

The datasets generated or analyzed during this study are not publicly available due to privacy concerns as the datasets may contain sensitive information of social media post authors.

Authors' Contributions

Conceptualization: ATC and LCW
Data curation: LCW and ATC
Formal analysis: LCW and KCP
Funding acquisition: ATC
Methodology: LCW and ATC

Software: LCW
Supervision: ATC
Validation: LCW and ATC
Visualization: LCW
Writing—original draft: LCW
Writing—review and editing: LCW, ATC, MC, and KCP

Conflicts of Interest

None declared.

Multimedia Appendix 1

Annotation guide for stigma mechanisms, settings, and actors.
[DOCX File , 3944 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Keywords for social ecological variables.
[DOCX File , 16 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Full results of the Kruskal-Wallis and Dunn post hoc tests.
[DOCX File , 144 KB-Multimedia Appendix 3]

References

1. Goffman E. Stigma: Notes on the Management of Spoiled Identity. New York City, NY. Touchstone; 2009.
2. Link BG, Phelan JC. Conceptualizing stigma. Annu Rev Sociol. 2001;27:363-385. [doi: 10.1146/annurev.soc.27.1.363]
3. Kulesza M, Ramsey S, Brown R, Larimer M. Stigma among individuals with substance use disorders: does it predict substance use, and does it diminish with treatment? J Addict Behav Ther Rehabil. Jan 15, 2014;3(1):1000115. [FREE Full text] [doi: 10.4172/2324-9005.1000115] [Medline: 25635257]
4. Garpenhag L, Dahlman D. Perceived healthcare stigma among patients in opioid substitution treatment: a qualitative study. Subst Abuse Treat Prev Policy. Oct 26, 2021;16(1):81. [FREE Full text] [doi: 10.1186/s13011-021-00417-3] [Medline: 34702338]
5. Baysan Arabaci L, Ayakdaş Dağli D, Taş G, Büyükbayram Arslan A. Stigmatization and social functioning levels of patients with alcohol use disorders. J Addict Nurs. 2020;31(4):295-301. [doi: 10.1097/JAN.0000000000000370] [Medline: 33264202]
6. Hathaway AD, Comeau NC, Erickson PG. Cannabis normalization and stigma: contemporary practices of moral regulation. Criminol Crim Justice. Aug 01, 2011;11(5):451-469. [doi: 10.1177/1748895811415345]
7. Charron E, Mayo RM, Heavner-Sullivan SF, Eichelberger KY, Dickes L, Truong KD, et al. "It's a very nuanced discussion with every woman": health care providers' communication practices during contraceptive counseling for patients with substance use disorders. Contraception. Nov 2020;102(5):349-355. [doi: 10.1016/j.contraception.2020.09.002] [Medline: 32941890]
8. Cernasev A, Hohmeier KC, Frederick K, Jasmin H, Gatwood J. A systematic literature review of patient perspectives of barriers and facilitators to access, adherence, stigma, and persistence to treatment for substance use disorder. Explor Res Clin Soc Pharm. Jun 04, 2021;2:100029. [FREE Full text] [doi: 10.1016/j.rcsop.2021.100029] [Medline: 35481114]
9. Kulesza M, Larimer ME, Rao D. Substance use related stigma: what we know and the way forward. J Addict Behav Ther Rehabil. May 27, 2013;2(2):782. [FREE Full text] [doi: 10.4172/2324-9005.1000106] [Medline: 25401117]
10. Earnshaw VA, Chaudoir SR. From conceptualizing to measuring HIV stigma: a review of HIV stigma mechanism measures. AIDS Behav. Dec 2009;13(6):1160-1177. [FREE Full text] [doi: 10.1007/s10461-009-9593-3] [Medline: 19636699]
11. Smith LR, Earnshaw VA, Copenhaver MM, Cunningham CO. Substance use stigma: reliability and validity of a theory-based scale for substance-using populations. Drug Alcohol Depend. May 01, 2016;162:34-43. [FREE Full text] [doi: 10.1016/j.drugalcdep.2016.02.019] [Medline: 26972790]
12. Chen AT, Johnny S, Conway M. Examining stigma relating to substance use and contextual factors in social media discussions. Drug Alcohol Depend Rep. May 05, 2022;3:100061. [FREE Full text] [doi: 10.1016/j.dadr.2022.100061] [Medline: 36845987]
13. Corrigan PW, Kerr A, Knudsen L. The stigma of mental illness: explanatory models and methods for change. Appl Prev Psychol. Sep 2005;11(3):179-190. [doi: 10.1016/j.appsy.2005.07.001]
14. Sudhinaraset M, Wigglesworth C, Takeuchi DT. Social and cultural contexts of alcohol use: influences in a social-ecological framework. Alcohol Res. 2016;38(1):35-45. [FREE Full text] [Medline: 27159810]

15. Ennett ST, Bauman KE, Hussong A, Faris R, Foshee VA, Cai L, et al. The peer context of adolescent substance use: findings from social network analysis. J Res Adolesc. Apr 25, 2006;16(2):159-186. [doi: 10.1111/j.1532-7795.2006.00127.x]
16. Glanz K, Rimer B, Viswanath K. Health Behavior: Theory, Research, and Practice. Hoboken, NJ. John Wiley & Sons; 2015.
17. Ingram L, Stafford C, Deming ME, Anderson JD, Robillard A, Li X. A systematic mixed studies review of the intersections of social-ecological factors and HIV stigma in people living with HIV in the U.S. South. J Assoc Nurses AIDS Care. 2019;30(3):330-343. [doi: 10.1097/JNC.0000000000000076] [Medline: 31021963]
18. Williams RS, Richards VL, Stetten NE, Canidate SS, Algarin A, Fiore A, et al. Applying the social ecological model to explore HIV-related stigma in Florida: a qualitative study. Stigma Health. Aug 2024;9(3):362-371. [FREE Full text] [doi: 10.1037/sah0000458] [Medline: 40134669]
19. Booth SL, Sallis JF, Ritenbaugh C, Hill JO, Birch LL, Frank LD, et al. Environmental and societal factors affect food choice and physical activity: rationale, influences, and leverage points. Nutr Rev. Mar 2001;59(3 Pt 2):S21-39; discussion S57-65. [doi: 10.1111/j.1753-4887.2001.tb06983.x] [Medline: 11330630]
20. van Kasteren YF, Lewis LK, Maeder A. Office-based physical activity: mapping a social ecological model approach against COM-B. BMC Public Health. Feb 03, 2020;20(1):163. [FREE Full text] [doi: 10.1186/s12889-020-8280-1] [Medline: 32013952]
21. Sallis JF, Cervero RB, Ascher W, Henderson KA, Kraft MK, Kerr J. An ecological approach to creating active living communities. Annu Rev Public Health. 2006;27:297-322. [doi: 10.1146/annurev.publhealth.27.021405.102100] [Medline: 16533119]
22. Yang G, King SG, Lin HM, Goldstein RZ. Emotional expression on social media support forums for substance cessation: observational study of text-based Reddit posts. J Med Internet Res. Jul 19, 2023;25:e45267. [FREE Full text] [doi: 10.2196/45267] [Medline: 37467010]
23. Medvedev AN, Lambiotte R, Delvenne JC. The anatomy of Reddit: an overview of academic research. In: Proceedings of the Dynamics On and Of Complex Networks III. 2017. Presented at: DOOCN 2017; June 19, 2017; Indianapolis, IN. [doi: 10.1007/978-3-030-14683-2_9]
24. Chi Y, Chen HY. Investigating substance use via Reddit: systematic scoping review. J Med Internet Res. Oct 25, 2023;25:e48905. [FREE Full text] [doi: 10.2196/48905] [Medline: 37878361]
25. Alambo A, Padhee S, Banerjee T, Thirunarayan K. COVID-19 and mental health/substance use disorders on Reddit: a longitudinal study. In: Proceedings of the Pattern Recognition. ICPR International Workshops and Challenges. 2021. Presented at: ICPR 2021; January 10-15, 2021; Virtual Event. [doi: 10.1007/978-3-030-68790-8_2]
26. Sowles SJ, Krauss MJ, Gebremedhn L, Cavazos-Rehg PA. "I feel like I've hit the bottom and have no idea what to do": supportive social networking on Reddit for individuals with a desire to quit cannabis use. Subst Abus. 2017;38(4):477-482. [FREE Full text] [doi: 10.1080/08897077.2017.1354956] [Medline: 28704167]
27. Liao M, Li Y, Kianifard F, Obi E, Arcona S. Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. BMC Nephrol. Mar 02, 2016;17:25. [FREE Full text] [doi: 10.1186/s12882-016-0238-2] [Medline: 26936756]
28. Oh W, Jayaraman P, Sawant AS, Chan L, Levin MA, Charney AW, et al. Using sequence clustering to identify clinically relevant subphenotypes in patients with COVID-19 admitted to the intensive care unit. J Am Med Inform Assoc. Jan 29, 2022;29(3):489-499. [FREE Full text] [doi: 10.1093/jamia/ocab252] [Medline: 35092685]
29. Burgel PR, Paillasseur JL, Caillaud D, Tillie-Leblond I, Chanez P, Escamilla R, et al. Clinical COPD phenotypes: a novel approach using principal component and cluster analyses. Eur Respir J. Sep 2010;36(3):531-539. [FREE Full text] [doi: 10.1183/09031936.00175109] [Medline: 20075045]
30. Haldar P, Pavord ID, Shaw DE, Berry MA, Thomas M, Brightling CE, et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med. Aug 01, 2008;178(3):218-224. [FREE Full text] [doi: 10.1164/rccm.200711-1754OC] [Medline: 18480428]
31. Baumgartner J, Zannettou S, Keegan B, Squire M, Blackburn J. The Pushshift Reddit dataset. Proc Int AAAI Conf Web Soc Media. May 26, 2020;14(1):830-839. [doi: 10.1609/icwsm.v14i1.7347]
32. MacLean D, Gupta S, Lembke A, Manning C, Heer J. 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. 2015. Presented at: CSCW '15; March 14-18, 2015; Vancouver, BC. [doi: 10.1145/2675133.2675146]
33. Roesler D, Johnny S, Conway M, Chen AT. A theory-informed deep learning approach to extracting and characterizing substance use-related stigma in social media. BMC Digit Health. Aug 16, 2024;2:60. [doi: 10.1186/s44247-024-00065-0]
34. Magwood O, Salvalaggio G, Beder M, Kendall C, Kpade V, Daghmach W, et al. The effectiveness of substance use interventions for homeless and vulnerably housed persons: a systematic review of systematic reviews on supervised consumption facilities, managed alcohol programs, and pharmacological agents for opioid use disorder. PLoS One. Jan 16, 2020;15(1):e0227298. [FREE Full text] [doi: 10.1371/journal.pone.0227298] [Medline: 31945092]
35. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. RoBERTa: a robustly optimized BERT pretraining approach. ArXiv. Preprint posted online on July 26, 2019. [doi: 10.48550/arXiv.1907.11692]

36. Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. ArXiv. Preprint posted online on October 11, 2018. [doi: 10.48550/arXiv.1810.04805]
37. Mohammad S. Word affect intensities. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. 2018. Presented at: LREC 2018; May 7-12, 2018; Miyazaki, Japan. URL: https://aclanthology.org/L18-1027 [doi: 10.1007/s10579-005-2692-5]
38. Strapparava C, Valitutti A. WordNet affect: an affective extension of WordNet. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation. 2004. Presented at: LREC'04; May 26-28, 2004; Lisbon, Portugal.
39. Pennebaker JW, Boyd RL, Jordan K, Blackburn K. The development and psychometric properties of LIWC2015. The University of Texas at Austin. 2015. URL: https://repositories.lib.utexas.edu/server/api/core/bitstreams/b0d26dcf-2391-4701-88d0-3cf50ebee697/content [accessed 2025-08-21]
40. Kumar V, Choudhary A, Cho E. Data augmentation using pre-trained transformer models. In: Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems. 2020. Presented at: LifeLongNLP 2020; December 7, 2020; Suzhou, China. [doi: 10.18653/v1/2020.lifelongnlp-1.0]
41. bert-base-uncased. Hugging Face. URL: https://huggingface.co/bert-base-uncased [accessed 2024-01-10]
42. roberta-base. Hugging Face. URL: https://huggingface.co/roberta-base [accessed 2024-01-10]
43. Bezdek JC, Ehrlich R, Full W. FCM: the fuzzy c-means clustering algorithm. Comput Geosci. 1984;10(2-3):191-203. [doi: 10.1016/0098-3004(84)90020-7]
44. Loftus TJ, Shickel B, Balch JA, Tighe PJ, Abbott KL, Fazzone B, et al. Phenotype clustering in health care: a narrative review for clinicians. Front Artif Intell. Aug 12, 2022;5:842306. [FREE Full text] [doi: 10.3389/frai.2022.842306] [Medline: 36034597]
45. Li X, Lu X, Tian J, Gao P, Kong H, Xu G. Application of fuzzy c-means clustering in data analysis of metabolomics. Anal Chem. Jun 01, 2009;81(11):4468-4475. [doi: 10.1021/ac900353t] [Medline: 19408956]
46. Assent I. Clustering high dimensional data. WIREs Data Min Knowl Discov. Jun 22, 2012;2(4):340-350. [doi: 10.1002/widm.1062]
47. Liu Y, Zhang X, Chen J, Chao H. A validity index for fuzzy clustering based on bipartite modularity. J Electr Comput Eng. Aug 08, 2019;2019:1-9. [doi: 10.1155/2019/2719617]
48. Ekemeyong Awong LE, Zielinska T. Comparative analysis of the clustering quality in self-organizing maps for human posture classification. Sensors (Basel). Sep 15, 2023;23(18):7925. [FREE Full text] [doi: 10.3390/s23187925] [Medline: 37765983]
49. Antonenko PD, Toy S, Niederhauser DS. Using cluster analysis for data mining in educational technology research. Educ Technol Res Dev. Feb 21, 2012;60:383-398. [doi: 10.1007/s11423-012-9235-8]
50. Bronfenbrenner U. The ecology of human developmentxperiments by nature and design. In: The Ecology of Human Development: Experiments by Nature and Design. Cambridge, MA. Harvard University Press; 1981.
51. Copeland M, Fisher JC, Moody J, Feinberg ME. Different kinds of lonely: dimensions of isolation and substance use in adolescence. J Youth Adolesc. Aug 2018;47(8):1755-1770. [FREE Full text] [doi: 10.1007/s10964-018-0860-3] [Medline: 29774451]
52. Can G, Tanrıverdi D. Social functioning and internalized stigma in individuals diagnosed with substance use disorder. Arch Psychiatr Nurs. Dec 2015;29(6):441-446. [doi: 10.1016/j.apnu.2015.07.008] [Medline: 26577560]
53. Ingram I, Kelly PJ, Deane FP, Baker AL, Goh MC, Raftery DK, et al. Loneliness among people with substance use problems: a narrative systematic review. Drug Alcohol Rev. Jul 2020;39(5):447-483. [doi: 10.1111/dar.13064] [Medline: 32314504]
54. Riffe D, Lacy S, Fico F, Watson B. Analyzing Media Messages: Using Quantitative Content Analysis in Research. New York, NY. Routledge; 2019.
55. Reif S, Braude L, Lyman DR, Dougherty RH, Daniels AS, Ghose SS, et al. Peer recovery support for individuals with substance use disorders: assessing the evidence. Psychiatr Serv. Jul 2014;65(7):853-861. [doi: 10.1176/appi.ps.201400047] [Medline: 24838535]
56. Eddie D, Hoffman L, Vilsaint C, Abry A, Bergman B, Hoeppner B, et al. Lived experience in new models of care for substance use disorder: a systematic review of peer recovery support services and recovery coaching. Front Psychol. Jun 13, 2019;10:1052. [FREE Full text] [doi: 10.3389/fpsyg.2019.01052] [Medline: 31263434]
57. Palamar JJ, Kiang MV, Halkitis PN. Predictors of stigmatization towards use of various illicit drugs among emerging adults. J Psychoactive Drugs. 2012;44(3):243-251. [FREE Full text] [doi: 10.1080/02791072.2012.703510] [Medline: 23061324]
58. McKight PE, Najab J. Kruskal-Wallis Test. In: The Corsini Encyclopedia of Psychology. Hoboken, NJ. John Wiley & Sons; 2010.
59. Ford E, Shepherd S, Jones K, Hassan L. Toward an ethical framework for the text mining of social media for health research: a systematic review. Front Digit Health. Jan 26, 2021;2:592237. [FREE Full text] [doi: 10.3389/fdgth.2020.592237] [Medline: 34713062]
60. Lingard L. Beyond the default colon: effective use of quotes in qualitative research. Perspect Med Educ. Dec 2019;8(6):360-364. [FREE Full text] [doi: 10.1007/s40037-019-00550-7] [Medline: 31758490]

61. van Boekel LC, Brouwers EP, van Weeghel J, Garretsen HF. Stigma among health professionals towards patients with substance use disorders and its consequences for healthcare delivery: systematic review. Drug Alcohol Depend. Jul 01, 2013;131(1-2):23-35. [doi: 10.1016/j.drugalcdep.2013.02.018] [Medline: 23490450]
62. Salamat S, Hegarty P, Patton R. Same clinic, different conceptions: drug users' and healthcare professionals' perceptions of how stigma may affect clinical care. J Appl Soc Psychol. May 24, 2019;49(8):534-545. [doi: 10.1111/jasp.12602]
63. Muncan B, Walters SM, Ezell J, Ompad DC. "They look at us like junkies": influences of drug use stigma on the healthcare engagement of people who inject drugs in New York city. Harm Reduct J. Jul 31, 2020;17(1):53. [FREE Full text] [doi: 10.1186/s12954-020-00399-8] [Medline: 32736624]
64. Quinn DM, Williams MK, Weisz BM. From discrimination to internalized mental illness stigma: the mediating roles of anticipated discrimination and anticipated stigma. Psychiatr Rehabil J. Jun 2015;38(2):103-108. [FREE Full text] [doi: 10.1037/prj0000136] [Medline: 25844910]
65. Corrigan PW, Rafacz J, Rüsch N. Examining a progressive model of self-stigma and its impact on people with serious mental illness. Psychiatry Res. Oct 30, 2011;189(3):339-343. [FREE Full text] [doi: 10.1016/j.psychres.2011.05.024] [Medline: 21715017]
66. Prizeman K, Weinstein N, McCabe C. Effects of mental health stigma on loneliness, social isolation, and relationships in young people with depression symptoms. BMC Psychiatry. Jul 21, 2023;23(1):527. [FREE Full text] [doi: 10.1186/s12888-023-04991-7] [Medline: 37479975]
67. Cho J, Goldenson NI, Kirkpatrick MG, Barrington-Trimis JL, Pang RD, Leventhal AM. Developmental patterns of tobacco product and cannabis use initiation in high school. Addiction. Feb 2021;116(2):382-393. [FREE Full text] [doi: 10.1111/add.15161] [Medline: 32533801]
68. Sampasa-Kanyinga H, Hamilton HA, LeBlanc AG, Chaput JP. Cannabis use among middle and high school students in Ontario: a school-based cross-sectional study. CMAJ Open. Jan 23, 2018;6(1):E50-E56. [FREE Full text] [doi: 10.9778/cmajo.20170159] [Medline: 29367264]
69. Kuntsche E, Jordan MD. Adolescent alcohol and cannabis use in relation to peer and school factors. Results of multilevel analyses. Drug Alcohol Depend. Sep 15, 2006;84(2):167-174. [doi: 10.1016/j.drugalcdep.2006.01.014] [Medline: 16542799]
70. Porath-Waller AJ, Beasley E, Beirness DJ. A meta-analytic review of school-based prevention for cannabis use. Health Educ Behav. Oct 2010;37(5):709-723. [doi: 10.1177/1090198110361315] [Medline: 20522782]
71. Weiss RD, Potter JS, Griffin ML, McHugh RK, Haller D, Jacobs P, et al. Reasons for opioid use among patients with dependence on prescription opioids: the role of chronic pain. J Subst Abuse Treat. Aug 2014;47(2):140-145. [FREE Full text] [doi: 10.1016/j.jsat.2014.03.004] [Medline: 24814051]
72. Schomerus G, Lucht M, Holzinger A, Matschinger H, Carta MG, Angermeyer MC. The stigma of alcohol dependence compared with other mental disorders: a review of population studies. Alcohol Alcohol. 2011;46(2):105-112. [doi: 10.1093/alcalc/agq089] [Medline: 21169612]
73. Skliamis K, Benschop A, Korf DJ. Cannabis users and stigma: a comparison of users from European countries with different cannabis policies. Eur J Criminol. Dec 23, 2020;19(6):1483-1500. [doi: 10.1177/1477370820983560]
74. Podina IR, Bucur AM, Todea D, Fodor L, Luca A, Dinu LP, et al. Mental health at different stages of cancer survival: a natural language processing study of Reddit posts. Front Psychol. Jun 23, 2023;14:1150227. [FREE Full text] [doi: 10.3389/fpsyg.2023.1150227] [Medline: 37425170]
75. Fry MS, Shircliff K, Benham M, Duncan T, Ladd K, Gilbert MK, et al. Medication assisted recovery: a social ecological approach to understanding how stigma shapes effective use. J Appl Soc Sci. Apr 11, 2023;17(2):220-240. [doi: 10.1177/19367244231159096]
76. Maina G, Marshall K, Sherstobitof J. Untangling the complexities of substance use initiation and recovery: client reflections on opioid use prevention and recovery from a social-ecological perspective. Subst Abuse. Oct 13, 2021;15:11782218211050372. [FREE Full text] [doi: 10.1177/11782218211050372] [Medline: 34675526]

Abbreviations

BERT: bidirectional encoder representations from transformers
FCM: fuzzy c-means
ML: machine learning
NLP: natural language processing
PCA: principal component analysis
RQ: research question
SEF: social ecological framework

Edited by A Mavragani; submitted 14.Nov.2024; peer-reviewed by K Berahmand, J Curtin; comments to author 20.Feb.2025; revised version received 26.May.2025; accepted 14.Aug.2025; published 13.Nov.2025

Please cite as:
Wang LC, Pike KC, Conway M, Chen AT
Identifying Stigma Phenotypes in Social Media Narratives of Substance Use: Observational Study
J Med Internet Res 2025;27:e68695
URL: https://www.jmir.org/2025/1/e68695
doi: 10.2196/68695
PMID: