Title:
Assessing the validity of online drug forums as a source for estimating demographic and temporal trends in drug use

Author list:
Michael J. Paul, Ph.D.[1*]
Margaret S. Chisolm, M.D.[2]
Matthew W. Johnson, Ph.D.[2]
Ryan G. Vandrey, Ph.D.[2]
Mark Dredze, Ph.D.[3]

[1] Department of Information Science, University of Colorado, Boulder, Colorado 80309, USA
[2] Department of Psychiatry and Behavioral Sciences, Johns Hopkins University, Baltimore, Maryland 21224, USA
[3] Department of Computer Science, Johns Hopkins University, Baltimore, Maryland 21218, USA

* Corresponding author

Abstract

Objectives: Addiction researchers have begun monitoring online forums to uncover self-reported details about use and effects of emerging drugs. The use of such online data sources has not been validated against data from large epidemiological surveys. This study aimed to characterize and compare the demographic and temporal trends associated with drug use as reported in online forums and in a large epidemiological survey.

Methods: Data were collected from the website, drugs-forum.com, from January 2007 through August 2012 (143,416 messages posted by 8,087 members) and from the United States National Survey on Drug Use and Health (NSDUH) from 2007-2012. Measures of forum participation levels were compared with and validated against two measures from the NSDUH survey data: percentage of people using the drug in last 30 days and percentage using the drug more than 100 times in the past year.

Results: For established drugs (e.g., cannabis), significant correlations were found across demographic groups between drugs-forum.com and the NSDUH survey data, while weaker, non-significant correlations were found with temporal trends. Emerging drugs (e.g., *Salvia divinorum*) were strongly associated with male users in the forum, in agreement with survey-derived data, and had temporal patterns that increased in synchrony with poison control reports.

Conclusions: These results offer the first assessment of online drug forums as a valid source for estimating demographic and temporal trends in drug use. The analyses suggest that online forums are a reliable source for estimation of demographic associations and early identification of emerging drugs, but a less reliable source for measurement of long-term temporal trends.

Introduction

Over the past two decades, the number of novel drugs entering the illicit drug market has expanded considerably. For example, in 2014, 101 unique and previously undetected psychoactive compounds were identified among drug substances confiscated and tested in Europe (European Monitoring Centre for Drugs and Drug Addiction, 2015). Among these, several drugs gained popularity and have become established entities in the illicit drug scene. Recent examples of "emerging drugs" include synthetic cathinones (aka "bath salts"), synthetic cannabinoids (e.g., "spice," "K2," "incense"), phenethylamines (e.g., "bromo dragonfly," 2C-chemicals), and *Salvia divinorum* (aka "salvia").

The proliferation of emerging drugs makes information on demographic and temporal trends in drug use critical for providing emergency and ongoing substance use treatment, developing clinically relevant research questions, creating effective public health campaigns, and crafting impactful public policy. The most common methods for obtaining this information are focus groups and surveys (Reyes et al. 2012; Hout & Bingham 2012), such as the National Survey on Drug Use and Health (NSDUH). While extremely valuable, NSDUH and similar surveys can fail to identify the emergence of new drugs (Dunn et al. 2011). For example, although some large surveys (e.g., Monitoring the Future) added items about synthetic cannabinoids and cathinone-derivative stimulants ("bath salts") in 2012, these drugs had emerged several years prior and there is still no population-level knowledge of prevalence or adverse effect rates of these drugs. This information gap means that clinicians, researchers, and policy makers may wait years before they can take meaningful action. Surveys remain essential, but there is a need for novel, faster methods to complement the existing annual population surveys.

Addiction researchers are turning increasingly to online sources (Corazza et al. 2011; Davey et al. 2012; Corazza et al. 2013) to uncover and disseminate details about use, effects, and popularity of a variety of emerging drugs (Morgan et al. 2010; Corazza et al. 2012; Gallagher et al. 2012). Comprehensive drug reviews now include these non-standard sources (Hill & Thomas 2011). In particular, online forums (web-based discussion communities) are popular avenues for discovering and sharing information about drug use (Wax 2002). For example, Drugs-Forum.com and BlueLight.com have been active for over a decade and contain 1 million and 4.4 million messages, respectively. These forums provide information regarding dose, preparation, and the type and duration of effects associated with use of various drugs or drug combinations. Online discussions on synthetic cannabinoids and "bath salts" date back to 2006, four years before they attracted clinical attention in the U.S. However, these types of online data sources consist entirely of a convenience sample of forum users, and the extent to which this information is representative of national drug use trends has not been examined. This study aims to characterize demographic and temporal trends as reported for various drugs in a large online drug forum, and compare these characteristics to nationally representative survey data to assess

the validity of online drug forums as a source for estimating demographic and temporal trends in drug use.

The dataset for this study was drawn from publicly available messages on drugs-forum.com. The forum contains multiple "subforums," each focused on a particular topic related to drug use, although most messages are associated with a subforum dedicated to a particular drug. These messages include self-reports of drug use experiences, including physical, pharmacological, and chemical characterizations of the drug and specifics regarding its use (e.g., dose; route of administration; context of use; type, magnitude, and duration of perceived effects). In addition, some members include public biographic profiles; all must indicate gender, and slightly more than half voluntarily provide age (55%) and country of residence (58%). There is no information about race or ethnicity.

The authors hypothesize that demographic and temporal trends from online drug forums and national survey data sources will align, suggesting the validity of online forums as a complementary source for estimating trends in drug use. If this hypothesis is supported, drug forums could be utilized as a source for discovering emerging trends in drug use in "real-time" using computer programs that track forum content. That data, in turn, could be used to inform relevant authorities (e.g., medical, public health, legislative, law enforcement) and to guide annual modifications to national epidemiological surveys such as NSDUH.

Methods

Data

Forum Data

The dataset was comprised of messages contained in 45 separate subforums on drugs-forum.com posted from January 2007 through August 2012. While other online communities exist, drugs-forum.com was chosen because, unlike other popular communities (e.g., BlueLight.com) it contains data on age, gender, and location of members; and content is organized into drug-specific subforums, enabling the analysis of drug-specific discussions.

Because NSDUH includes only data from respondents living in the US, for this study, the forum dataset was limited to include only data from forum members publicly claiming to be living in the US (57% of members). Thus limited, the forum dataset contains 143,416 messages by 8,087 members. More details about the data are provided in the supplement. These data were originally collected for a study summarizing the characteristics of emerging drugs from online content (Paul & Dredze 2013).

This study and use of data was judged to be exempt by the Johns Hopkins University Institutional Review Board.

Survey Data

The dataset was comprised of survey data from the US National Survey on Drug Use and Health (NSDUH). NSDUH is a large annual nationwide survey conducted through approximately 70,000 interviews of individuals aged 12 and older. The present analysis only considered data for ages 18 and older, because members of drugs-forum.com must endorse an age of 18+ years to participate. More details of the NSDUH methodology are available at: https://nsduhweb.rti.org/respweb/homepage.cfm.

The dataset was downloaded from the Drug Abuse and Mental Health Data Archive (2014), which provides aggregate weighted prevalence estimates of various drugs, grouped into broad categories (described in Table 1, along with the corresponding website subforums). Yearly prevalence is provided for male and female genders and the following age groups: 18-25, 26-34, 35-49, 50+ years. The study used weighted data. NSDUH data are weighted to adjust for unequal sample selection probabilities throughout the survey process.

Survey Comparison and Validation

Two measures of forum participation were computed. First, the percentage of forum members who posted in one of the relevant subforums was computed (Table 1, second column). For demographic trends, the number of members of a gender or age group who posted at least one message in a particular drug's subforum was computed, and then that number was divided by the number of all members of that gender or age group. For temporal trends, the number of members who posted at least one message in a drug's subforum in a particular year was computed, and then that number was divided by the number of all members who wrote at least one message in that year.

Second, the percentage of forum members who wrote a message containing a relevant keyword was computed (Table 1, third column). This metric is determined similarly to the first, but using the criterion that a member must have written at least one message containing a drug-related keyword in any subforum, rather than have simply posted a message in a drug's subforum. The keywords for each drug category include the full names of all drugs and drug classes specified by the forum, augmented with a small set of slang terms and paraphernalia terms commonly observed in the forum, using a word association technique described in the supplement. The list was curated by one researcher and independently corroborated as valid by addiction clinicians and researchers on the study team.

Drug category	**Drugs-Forum.com subforum names**	**Keywords used for filtering**
Established Drugs		
Marijuana	*Cannabis* and its subforums	marijuana, cannabis, weed, pot, grass, bud, joint, blunt, ganja, MJ
Cocaine	*Cocaine & Crack*	cocaine, coke, crack
Hallucinogens (LSD, psilocybin, Peyote and mescaline, MDMA, PCP)	*LSD*, *Peyote & San Pedro*, *Ecstasy & MDMA* (subforums pertaining to psilocybin were not publicly accessible)	ecstasy, MDMA, LSD, acid, mushrooms, shrooms, psilocybin, Peyote, San Pedro, mescaline
Pain Relievers (nonmedical use of various opioids)	All subforums categorized under *Opiates & Opioids*, excluding those not included in this NSDUH category: *Heroin*, *Morphine*, *Opium & Poppy*, *Buprenorphine*, *Fentanyl*	pain( )killer(s), opioid(s), opiate(s), codeine, hydrocodone, hydromorphone, methadone, morphine, oxycodone, oxymorphone, tramadol, Vicodin, Oxycontin, Percocet
Stimulants (including amphetamines)	*Amphetamine*, *Adderall*, *Concerta & Ritalin*, *Methamphetamine*	stimulants, amphetamine(s), speed, Adderall, Concerta, Ritalin, meth, methamphetamine
Tranquilizers (including benzodiazepines)	*Downers and sleeping pills*, *Benzodiazepines*	downers, sleeping pills, Ambien, benzo(s), benzodiazepine(s)
Emerging Drugs		
Synthetic cathinones (e.g. “bath salts”)	*Beta-Ketones*	bath salt(s), (beta-)ketone(s), mephedrone, MDPV
Synthetic cannabinoids (e.g. “spice” and “K2”)	*Cannabinoids*	cannabinoid(s), spice, K2, JWH, JWH-018, JWH-073, JWH-200
Phenethylamines	*Phenethylamines*	phenethylamine(s), 2C(s), 2C-B, 2C-E, 2C-I, dragonfly
Salvia divinorum	*Salvia divinorum*	salvia

Table 1: The six established drug subcategories (defined by the survey data set) and four emerging drug subcategories, along with the corresponding subforums grouped for experimental comparison, and the list of keywords used to measure discussion of the corresponding drug category.

Both metrics measure interest in a drug, either by participation in that drug's designated forum or written reference to the drug. Certainly, interest is not the same as use; for example, someone might discuss a drug to advocate against its use. However, these metrics are straightforward to compute and broadly characterize the level of participation in the forums, which are hypothesized to correlate with drug use for the purpose of this study.

This study compares these two forum metrics (posting in a subforum or posting a keyword) with two usage items from the NSDUH survey data: recent use (the percentage of respondents who reported using a particular drug within the past 30 days) and frequent use (the percentage who reported using a drug more than 100 times in the past year), for each demographic group and time period.

Analyzing Emerging Drug Trends

In this study, drugs are categorized as either "established drugs" or "emerging drugs." "Established drugs" are defined as drugs for which non-medical use was well documented prior to the years in which data was collected for this study, demonstrated by their inclusion in NSDUH. "Established drugs" include: cannabis (marijuana); cocaine; hallucinogens included in NSDUH, including LSD and MDMA; opioids; stimulants, including amphetamines; and tranquilizers, including benzodiazepines. "Emerging drugs" are defined as drugs introduced to the illicit drug market or for which a notable expansion of use coincided with the timeframe of the data sources (2007-2012). "Emerging drugs" during this time period included: synthetic cathinones ("bath salts") including mephedrone and MDPV; synthetic cannabinoids (products commonly referred to as "incense," "spice," or "K2," among other brand names); phenethylamines (which includes "bromo dragonfly" and 2C- chemicals); and *Salvia divinorum* ("salvia"). The list of relevant subforums and keywords are provided in Table 1. Of these four "emerging drugs" subcategories, only *Salvia divinorum* was included in the NSDUH survey during the time interval covered by the study's dataset.

Modeling Demographic and Temporal Drug Associations

Associations between drugs and demographic groups were measured using a linear regression model (specified in the supplement). For each drug and demographic group, the prevalence is modeled with a variable for the drug, a variable for the demographic group, and an interaction term for the pairing of the drug with the demographic group. The interaction term signifies a positive or negative association between each demographic group and each drug. A positive (or negative) association means that a particular demographic group is more (or less) likely to use a certain drug than would be expected based on how likely the demographic group uses drugs in general (regardless of the specific drug) and how popular the drug is (regardless of the specific demographic group). The model can therefore account for differences in forum participation

levels, and -- even though the forum population is not representative of the survey population -- the model can still provide demographic associations after controlling for these differences.

To compare the survey and forum data, the model was estimated for the six established drug categories shown in Table 1, and then separately modeled for the four emerging drugs. The same linear regression model was used for estimating temporal associations with each drug and each year. While more complex models could be considered to understand interactions between the various attributes (gender, age, year), or to consider the effects of potential confounds, this would not directly address the main research question, which is to compare the forum and survey trends, rather than to explain the trends. As such, the study used a model of minimal complexity to improve interpretability and reduce chance of type I errors.

Results

Demographic Composition

Within the US, 82.9% of forum members were male. In contrast, 63.3% (weighted) of the NSDUH survey respondents who reported past month drug use (pooled from 2007-2012) were male.

The distribution of age groups for the forum and survey data were very similar. 54.8% of the forum members were ages 18-25 years (54.8% of the NSDUH past month use population, weighted), 27.9% of members were ages 26-34 years (31.5% of NSDUH, weighted), 13.0% of members were ages 35-49 years (10.5% of NSDUH, weighted), and 4.2% of members were ages 50+ years (3.2% of NSDUH, weighted).

Comparison of Established Drugs Data between Forums and Survey

Demographic Trends

Both the survey and forum models suggest greater drug use among male users and younger users. For demographic associations with specific drugs, Figure 1 shows scatter plots comparing the survey and forum demographic associations. The scatter plots show the values of the interaction terms in the regression models, which quantify the association between each drug and each demographic group. To quantify the relationship between the forum-derived associations and survey-derived associations, the Pearson correlation coefficient (r) was calculated between the forum association values and survey association values.

Table 2 shows that demographic trends derived from the survey and forum data are well correlated, with a median correlation of .565 (95% CI: -0.014–0.860, p=.056) and .492 (95% CI:

0.110–0.747, p=.015) for gender and age, respectively. These correlations are considered strong effects (Cohen 1988). Comparing the forum to the survey, forum participation was found to be a stronger indicator of frequent use of a drug rather than recent use, for both gender and age.

Table 3 shows ratios of the drug associations between demographic groups, using the best combination of metrics according to Table 2. For example, a gender ratio of 1.55 for cannabis means that males are 1.55 times more likely to use cannabis than females, relative to how often males use drugs compared to females in general. For brevity, only ratios between two groups (the two genders and the two most prominent age groups in the forum data, 18-25 years and 26-34 years) were computed, but full tables of all regression coefficients, along with detailed explanations of the calculation of these ratios, are provided in the supplement.

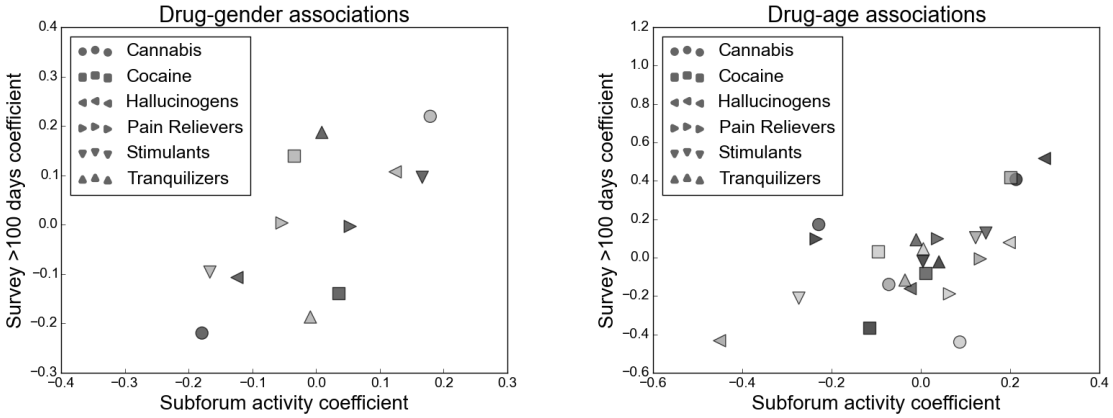

Figure 1: Scatter plots of the drug association values for gender (left; r = .659, p = .020) and age (right; r = .578, p = .003), as estimated with the survey data (y-axis) and forum data (x-axis). The circle shape indicates the drug (given in the legend) while the shading indicates the demographic attribute. Darker points are female coefficients in the gender plot, and darker points correspond to younger age groups in the age plot.

The two data sources (forum and NSDUH survey) showed very similar associations between gender and use of cannabis, hallucinogens, stimulants, and pain relievers. Both data sources found that cannabis and hallucinogens were more associated with males, stimulants were more associated with females, and pain relievers were about equally likely to be associated with males and females. However, differences between the two data sources were found when examining associations with cocaine and tranquilizers. Cocaine was much more likely to be associated with males in the NSDUH survey data, but cocaine was slightly more associated with females in the forum data. Tranquilizers were more likely to be associated with females in the survey data, but were balanced between the genders in the forum data.

	Gender		Age		Year	
Survey / Forum	Subforum	Keywords	Subforum	Keywords	Subforum	Keywords
Past month	.568 (p=.054)	.476 (p=.117)	.487 (p=.016)	.327 (p=.118)	.088 (p=.610)	.273 (p=.108)
>100 days	.659 (p=.020)	.562 (p=.057)	.578 (p=.003)	.496 (p=.014)	.059 (p=.732)	.247 (p=.147)

Table 2: Pearson correlation coefficients between the demographic and temporal association coefficients for the various survey metrics and forum metrics.

Drug	**Source**	**Gender Ratio (Male/Female)**	**Age Ratio (18-25/26-34)**	**Time Ratio (2012/2007)**
Established Drugs				
Cannabis	Survey	1.55	1.26	1.25
	Forum	1.43	1.56	0.97
Cocaine	Survey	1.39	0.75	0.70
	Forum	0.93	0.88	0.76
Hallucinogens	Survey	1.24	1.97	1.00
	Forum	1.28	1.35	0.70
Pain Relievers	Survey	1.01	1.00	0.91
	Forum	0.90	0.76	1.14
Stimulants	Survey	0.83	0.86	1.00
	Forum	0.72	0.87	1.15
Tranquilizers	Survey	0.69	0.89	1.25
	Forum	0.98	1.05	1.07
Emerging Drugs				
Synth. Cathinones	Forum	1.52	0.59	2.52
Synth. Cannabinoids	Forum	2.60	0.71	2.31
Phenethylamines	Forum	2.09	1.13	0.34
Salvia divinorum	Forum	1.60	1.35	0.23

Table 3: The ratios of drug associations for different demographic groups or time periods. For gender, age, and time, a ratio >1 means that the drug is more likely to be used by males than females, to be used by the youngest age group than the next oldest age group, or to be used in the most recent year of data than the oldest year of data, respectively; a ratio <1 means the inverse.

When examining associations between age and drug use, the associations were consistent between both data sources (forum and NSDUH survey) for all drugs except tranquilizers. Cannabis and hallucinogens were most associated with the 18-25 years age group in both data sources, stimulants were most associated with the 26-34 years age group in both sources, and cocaine and pain relievers were most associated with the 35-49 years age group in both sources. The tranquilizers subcategory did not show a consistent age trend across the survey and forum.

Temporal Trends

Unlike the demographic trends, the temporal trends are not significantly correlated with the survey (Table 2). Most of the drug categories did not have strong and consistent trends over the six years, and most of the year-specific associations were not statistically significant.

The largest trend agreement is with cocaine, which shows a fairly steady decrease over the six years across all four metrics. The largest mismatch is with cannabis, which increases heavily and nearly monotonically in the survey data, but decreases in the forum data. The decline over time is stark under the subforum activity metric, though there is only small temporal variation under the keyword metric, which suggests that forum members are making references to cannabis somewhat consistently over time, even though they are no longer posting in the dedicated subforums.

Analogous to the demographic association ratios, Table 3 shows ratios comparing the temporal associations for the most recent year of data (2012) to the earliest year of data (2007).

Emerging Drug Associations

Demographic Trends

All four emerging drugs are much more heavily associated with male forum members as compared to the established drugs. Men are especially likely to use synthetic cannabinoids and phenethylamines relative to women, according to the forum-derived results. These results in agreement with the NSDUH data on *Salvia divinorum*, which shows much higher use by males. An association with male gender has also been found in other studies on the use of synthetic cathinones (Johnson & Johnson 2014; Winstock et al. 2011), synthetic cannabinoids (Vandrey et al. 2012), and phenethylamines (Lawn et al. 2014).

There was not a consistent shift in age associations when comparing the emerging and established drugs. Among age groups, synthetic cathinones are most associated with the 26-34 years age group, synthetic cannabinoids are increasingly associated with older age groups, and

both phenethylamines and *Salvia divinorum* have a heavy association with the youngest group, 18-25 years.

Temporal Trends

The yearly forum activity for synthetic cathinones is extremely low prior to 2009, when the forum discussion begins to increase, peaking in 2010 (subforum metric) and 2011 (keyword metric). This closely matches the yearly number of "bath salts"-related reports to US poison control centers, which first appeared in 2010 and peaked in 2011. Data reported to the National Forensic Laboratory Information System (NFLIS) follow a similar trend, but first appear in 2009.[123] A number of US states outlawed the common synthetic cathinones in 2011, which may explain the drop in forum interest after this time point.[4]

The forum activity for synthetic cannabinoids is also extremely low prior to 2009 and is most prominent in 2010-2012, peaking in 2010 under both metrics. Poison control and NFLIS reports for synthetic cannabinoids have similar trends as synthetic cathinones, first appearing in 2009-2010, increasing rapidly, and peaking in 2011.[5]

The forum activity for phenethylamines decreases fairly consistently over time, with peaks in 2007. This long-term trend contrasts with NFLIS data, which shows reports increasing steadily from 2006 to 2010, although the peak of forum activity (2007-2008) coincides with the first years that had a substantial number of reports (after 2006, which had few reports). [6] A number of US states outlawed common phenethylamines in 2011, which may explain the drop in forum interest.[78]

The forum activity for *Salvia divinorum* decreases over time, with the highest temporal variability of any established or emerging drug. During this timeframe, several US states passed legislation restricting this drug.

---

[1] https://aapcc.s3.amazonaws.com/files/library/Bath_Salts_Web_Data_through_9.2014.pdf
[2] http://www.justice.gov/archive/ndic/pubs44/44571/44571p.pdf
[3] http://www.justice.gov/dea/resource-center/DIR-017-13%20NDTA%20Summary%20final.pdf
[4] https://www.erowid.org/chemicals/mdpv/mdpv_law.shtml
[5] http://www.justice.gov/dea/resource-center/DIR-017-13%20NDTA%20Summary%20final.pdf
[6] http://www.deadiversion.usdoj.gov/nflis/spec_rpt_emerging_2012.pdf
[7] https://www.erowid.org/chemicals/2ci/2ci_law.shtml
[8] https://www.erowid.org/chemicals/2ce/2ce_law.shtml

Discussion

The study's analyses showed that statistically significant correlations exist between drugs-forum.com and the NSDUH survey, when measuring associations between six established drug subcategories and age and gender groups. The age distributions in the two datasets are extremely similar, and the gender distributions in both are skewed toward males. The two established drug subcategories that did not align well with gender associations – cocaine and tranquilizers – have much fewer forum messages than the other four established drug subcategories, a possible cause of divergent findings. Excluding these two outlier drug subcategories, the median correlation coefficient between forum and survey gender associations was .960 ($p<.001$).

There was a weaker correlation with temporal trends. This is likely in part a consequence of established drugs showing relatively little variation in use and interest over the six years of data, as most of the drugs did not have a strong temporal trend in a consistent direction. Another explanation is that online drug forums are used in particular ways that do not align with general prevalence. For example, cannabis-related discussion declined over the six years in the forum, yet cannabis use increased over the same period according to the NSDUH survey data. One possible explanation for the discrepancy is that as cannabis use becomes more commonplace, it warrants less discussion, even if the members may still be cannabis users. This hypothesis is supported by the result that the trend declines less heavily when measuring keyword-based references to the drug (in any subforum) as compared to measuring member posts in the specific cannabis-related subforums.

The weak correlation with the overall temporal trends suggests that the forum data would not be amenable to longitudinal studies. In general, social media data are challenging for longitudinal research because people may use social media inconsistently over time. However, despite this limitation, the results suggest that online forums are still promising as a source for early detection of rising interest in emerging drugs. The results show that forum-derived temporal trends of emerging drugs rise in synchrony with rises in poison control reports. These results suggest it would be possible to develop an automated system to identify rising popularity of emerging drugs. Such technology could be utilized in parallel with poison control and emergency department data to alert medical, public health, legislative, and law enforcement officials about the identity of novel drugs coming to the market as well as detailed information on drug use practices reported by users on the forums.

One limitation of this study is that only one online drug forum website was analyzed. This particular data source was chosen because of its relative popularity and organized structure – drug-specific subforums and member profiles – which allowed examination of associations between subforum activity and self-reported demographic attributes. Even still, the forum-derived measurements do not have perfect accuracy and are only proxies for drug use. Members

may post to a subforum to ask questions even if they have not personally used the drug, and likewise when writing drug-related keywords. A potential solution for future research is the use of *natural language processing* (NLP), a computer science discipline that automates language understanding. For example, NLP tools have been shown to improve tracking of temporal health trends in social media (Abbasi et al. 2014), including drug-related social media to identify adverse drug reactions (ADRs) (Chary et al. 2013, Yates et al. 2013, Sarker et al. 2015). However, it is unknown whether and how much NLP would improve this task, and thus this study used simpler methods as an important starting point for analysis.

Another important study limitation is that the forum does not constitute a representative sample of the general population or even the population of people who use the internet. Rather, this is a community of people who already have an active interest in drugs, which poses challenges for estimating population-level prevalence. This mismatch may in part contribute to the poor correlation for temporal trends. General-purpose social media websites, such as Facebook and Twitter, may better serve this purpose, although these websites do not typically provide anonymity, and thus members may be reluctant to publicly discuss drug-related activities. Recent studies have shown that tobacco (Cobb et al. 2011), alcohol (Moreno et al. 2012), and prescription pain reliever use (Hanson et al. 2013) are discussed on Twitter. To the best of the authors' knowledge, this study is the first to validate social media trends for the wide range of drugs included here.

Finally, this report emphasizes that online trends need not be perfect to have utility. Traditional large surveys, such as NSDUH, will remain the gold standard, particularly for long-term trends. However, survey results are only published annually, with a lag of 1-2 years. Moreover, emerging drugs, by nature, cannot be included in these surveys until sometime (typically several years) after initial use. These characteristics make such surveys unsuitable for timely identification of emerging drugs and for monitoring fine-grained trends related to the rapid rise of a newly introduced drug. Online communities and social media data offer a way around these limitations, and can complement traditional surveys with large-scale, real-time reports. This study offers evidence that online forums can be used to estimate demographic associations with drugs, which can help identify at-risk subpopulations. The results also suggest that forum data can help detect rises in interest in emerging drugs, which coincide with increased poison control reports. Thus, the methods presented here may be used to identify and describe actionable patterns surrounding novel drugs, and to serve as a source of information for adding emerging drugs to large surveys such as NSDUH. By identifying novel drugs early, and by estimating how rapidly interest in a drug is rising and within which demographic groups, the proposed analyses of forum data can be leveraged to inform early interventions to prevent overdose deaths.

References

Abbasi, A. et al., 2014. Social Media Analytics for Smart Health. *IEEE Intelligent Systems*, 29(2), pp.60–80.

Chary M., Genes N., McKenzie A., Manini A. F., 2013. Leveraging social networks for toxicovigilance. *J Med Toxicol*, 9, pp.184–191.

Cobb, N.K. et al., 2011. Online Social Networks and Smoking Cessation: A Scientific Research Agenda. *Journal of Medical Internet Research*, 13(4).

Cohen J., 1988. *Statistical Power Analysis for the Behavioral Sciences*. L. Erlbaum Associates.

Corazza, O. et al., 2011. Designer drugs on the Internet: a phenomenon out-of-control? The emergence of hallucinogenic drug Bromo-Dragonfly. *Current Clinical Pharmacology*, 6(2), pp.125–129.

Corazza, O. et al., 2012. Phenomenon of new drugs on the Internet: the case of ketamine derivative methoxetamine. *Human Psychopharmacology: Clinical and Experimental*, 27(2), pp.145–149. Available at: http://dx.doi.org/10.1002/hup.1242.

Corazza, O. et al., 2013. Promoting innovation and excellence to face the rapid diffusion of novel Psychoactive substances in the EU: The outcomes of the reDNet project. In *Human Psychopharmacology*. pp. 317–323.

Davey, Z. et al., 2012. e-Psychonauts: Conducting research in online drug forum communities. *Journal of Mental Health*, 21(4), pp.386–394.

Drug Abuse and Mental Health Data Archive, 2014. Available at: http://www.icpsr.umich.edu/icpsrweb/SAMHDA/browse [Accessed June 19, 2014].

Dunn, M. et al., 2011. Effectiveness of and challenges faced by surveillance systems. *Drug Testing and Analysis*, 3(9), pp.635–641. Available at: http://dx.doi.org/10.1002/dta.333.

European Monitoring Centre for Drugs and Drug Addiction, 2014. Available at: http://www.emcdda.europa.eu/system/files/publications/974/TDAT15001ENN.pdf

Gallagher, C.T. et al., 2012. 5,6-Methylenedioxy-2-aminoindane: from laboratory curiosity to `legal high'. *Human Psychopharmacology: Clinical and Experimental*, 27(2), pp.106–112. Available at: http://dx.doi.org/10.1002/hup.1255.

Hanson, L.C. et al., 2013. An Exploration of Social Circles and Prescription Drug Abuse Through Twitter. *J Med Internet Res*, 15(9), p.e189. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24014109.

Hill, S.L. & Thomas, S.H.L., 2011. Clinical toxicology of newer recreational drugs. *Clinical Toxicology*, 49(8), pp.705–719. Available at: http://informahealthcare.com/doi/abs/10.3109/15563650.2011.615318.

Hout, M.C. Van & Bingham, T., 2012. Costly Turn On: Patterns of use and perceived consequences of mephedrone based head shop products amongst Irish injectors. *International Journal of Drug Policy*.

Johnson, P.S. & Johnson, M.W., 2014. Investigation of "bath salts" use patterns within an online sample of users in the United States. *J Psychoactive Drugs*, 46(5), pp.369–78.

Lawn, W. et al., 2014. The NBOMe hallucinogenic drug series: Patterns of use, characteristics of users and self-reported effects in a large international sample. *Journal of psychopharmacology (Oxford, England)*, 28(8), pp.780–788. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24569095.

Moreno, M.A. et al., 2012. Associations Between Displayed Alcohol References on Facebook and Problem Drinking Among College Students. *Archives of Pediatrics and Adolescent Medicine*, 166(2), pp.157–163.

Morgan, E.M., Snelson, C. & Elison-Bowers, P., 2010. Image and video disclosure of substance use on social media websites. *Computers in Human Behavior*, 26(6), pp.1405–1411. Available at: http://www.sciencedirect.com/science/article/pii/S0747563210000932.

Paul, M. & Dredze, M., 2013. Drug Extraction from the Web: Summarizing Drug Experiences with Multi-Dimensional Topic Models. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Reyes, J. et al., 2012. The Emerging of Xylazine as a New Drug of Abuse and its Health Consequences among Drug Users in Puerto Rico. *Journal of Urban Health*, pp.1–8.

Sarker, A., Ginn, R., Nikfarjam, A., et al., 2015. Utilizing social media data for pharmacovigilance: a review. *Journal of Biomedical Informatics*. 54, pp.202–212.

Vandrey, R. et al., 2012. A survey study to characterize use of Spice products (synthetic cannabinoids). *Drug and Alcohol Dependence*, 120(1-3), pp.238–241.

Wax, P.M., 2002. Just a click away: recreational drug web sites on the internet. *Pediatrics*, 109(6), pp.e96–e96.

Winstock, A. et al., 2011. Mephedrone: Use, subjective effects and health risks. *Addiction*, 106(11), pp.1991–1996.

Yates, A., Goharia, N., 2013. ADRTrace: Detecting Expected and Unexpected Adverse Drug Reactions from User Reviews on Social Media Sites. *Proceedings of the 35th European Conference on Advances in Information Retrieval*.

Supplement

Survey Comparison

There are some drug categories in the NSDUH survey that are not well represented in the forum and so these are not analyzed in the study. For example, the "Inhalants" category in the NSDUH includes a variety of pharmacologically unrelated drugs including volatile anesthetics (e.g., glue and gasoline) and nitrous oxide. Yet, only nitrous oxide has a category in drugs-forums.com. Thus it was decided to exclude the "Inhalants" category from the analysis because of the very small amount of data: fewer than one fifth the number of messages of the next smallest drug category, tranquilizers. NSDUH also contains a "Sedatives" category, which includes barbiturates, but no barbiturates are included in the forum. Benzodiazepines are categorized separately as "Tranquilizers" in NSDUH.

Data Statistics

In total, the dataset contains 351,787 messages posted by 24,424 members. Of the members who self-reported their location, 57% were from the US, with the United Kingdom being the next highest at 16%. For the study, the dataset was restricted to US members. The US data statistics are broken down by year in Table S1.

Table S2 shows the number of members and messages in study US dataset broken down by each drug category included in study experiments.

Table S3 shows the number of members and messages in study US dataset broken down by each gender and age. The member age is measured at the time of the member's earliest forum activity.

Demographic Composition

The NSDUH survey data gives the proportion of respondents with a demographic attribute who had used drugs, whereas the distribution this study reports is the proportion of drug-using respondents with each attribute. The former is converted into the latter using Bayes' theorem, by multiplying the prevalence rate by the true population distribution for the demographic attributes (as given by the 2010 US Census), and then re-normalizing.

Year	# members	# messages
2007	560	13,819
2008	791	18,125
2009	1,445	27,077
2010	2,491	29,407
2011	3,314	30,070
2012	3,006	24,918

Table S1. The number of active members and forum messages in each year of study dataset, among members from the US. Numbers in 2012 exclude activity after August, when the data was collected.

Drug category	# members	# messages
Established Drugs		
Cannabis	1,482	10,496
Cocaine	1,125	5,951
Hallucinogens	1,584	12,511
Pain relievers	2,369	20,053
Stimulants	2,571	21,010
Tranquilizers	930	5,192
Emerging Drugs		
Synthetic cathinones	597	2,982
Synthetic cannabinoids	1,035	6,768
Phenethylamines	475	2,224
Salvia divinorum	426	2,438

Table S2. The number of active members and forum messages from the US in the subforums of each drug category.

Group	# members	# messages
Female	1,383	21,512
Male	6,704	121,904
18-25	2,426	45,341
26-34	1,236	20,066
35-49	576	8,584
50+	185	2,328

Table S3. The number of active members and forum messages from the US within each gender and age group.

Demographic Associations

Methods

Associations between drugs and demographic groups were modeled with the following linear regression model. The log of the prevalence rate y_{dg} of using a drug d among those with gender g was modeled as:

(1) $\log y_{dg} = \beta_0 + \beta_d + \beta_g + \beta_{dg}$

And similarly the log of the prevalence rate for drug d among those in age group a was modeled as:

(2) $\log y_{da} = \beta_0 + \beta_d + \beta_a + \beta_{da}$

β_0 is the intercept, β_d is a drug-specific intercept, β_g and β_a are gender- and age-specific intercepts, and the β_{dg} and β_{da} variables are interaction coefficients between specific drug-gender or drug-age pairs.

The β_0 intercept captures the overall degree of drug use under the given metric, independent of any specific drug or demographic group. The β_d intercepts adjust for overall prevalence of each specific drug, and the β_g and β_a intercepts adjust for overall tendencies of each group to use drugs. The β_{dg} and β_{da} interaction variables can then be interpreted as associations or preferences between each drug and each demographic group. For example, if $\beta_{d\mathrm{M}}$ is higher than $\beta_{d\mathrm{F}}$ for a drug d, this means that men are more likely than women to use the drug d, after controlling for the overall likelihood of men and women using any drug.

By regressing against the log of the values, the coefficients can be interpreted as multiplicative rather than additive terms, so that the model captures *relative* differences in drug use between the demographic groups. This is important because the absolute differences vary dramatically by drug. Because the relative difference is not well defined when one value is zero, 0.1 is added to all y values (the lowest non-zero value in the survey data). This is standard technique in machine learning known as "smoothing" to avoid values of 0 in probabilistic models.

To compare the survey and forum data, the model was estimated for the six established drugs, and separately estimated for the four emerging drugs. The drug-gender and drug-age associations between the survey and forum data were directly compared by examining the coefficients estimated from the two datasets. The y values for the forum data were computed across the entire dataset, independent of year. Since the NSDUH survey data is given per year, the y values are averaged across the six years.

Results: Established Drugs

Table S4 shows the regression model coefficients for drug associations with gender and age, regressed against the two survey metrics (past month use and frequent use) and the two forum metrics (subforum activity and keyword activity), for the six established drugs. The table shows the intercept and the demographic-specific term, overall (β_0 and β_g, β_a) and for each specific drug category (β_d and β_{dg}, β_{da}).

To assess statistical significance of each drug-specific demographic association (interaction variables β_{dg}, β_{da}), the *r*-squared of the regression models were compared with and without each variable. If excluding the interaction variable reduces the *r*-squared significantly with $p<.05$, then the coefficient is considered significant. This means that the interaction of the drug and demographic group contains significant information beyond the overall drug and demographic intercepts (β_d and β_g, β_a).

For the scatter plots in Figure 1, each drug's coefficients are centered around zero by subtracting the mean of the two coefficients, to provide a more direct and interpretable comparison of the two data sources.

The ratios in Table 3 are calculated as follows. Because the regression model uses the log of the values, the regression coefficients were first exponentiated so that they are interpreted as probabilities rather than log-probabilities. Then the ratio of the exponentiated coefficients for the two demographic groups for each drug were computed. For example, the male-female ratio for a drug *d* is calculated as:

(3) $\exp(\beta_{d\mathrm{M}}) / \exp(\beta_{d\mathrm{F}}) = \exp(\beta_{d\mathrm{M}} - \beta_{d\mathrm{F}})$.

Results: Emerging Drugs

Table S5 shows the demographic regression model coefficients for the four emerging drugs.

To calculate the ratios in Table 3, additional adjustments were made to account for the fact that the overall demographic shift for emerging drugs may differ from established drugs. Because the regression model for emerging drugs is computed on only the four emerging drugs, rather than all drugs in the data, the drug-specific demographic associations are relative to the four emerging drugs, rather than the six established drugs. For example, all four emerging drugs are more likely to be used by males than the established drugs, which is captured by the regression model with a much higher value of the demographic-specific intercept, β_g or β_a. However, some of the drug-specific interaction variables, β_{dg} or β_{da}, have a higher female coefficient since they have a higher

female association relative to the other emerging drugs, even if they have a lower female association relative to the six established drugs. To adjust for this, the log-difference between the forum demographic coefficients and the survey demographic coefficients were included. For example, the gender ratio for a drug *d* is calculated as:

(4) $$\exp((\beta_{\mathrm{M,forum}} - \beta_{\mathrm{M,survey}}) - (\beta_{\mathrm{F,forum}} - \beta_{\mathrm{F,survey}}) + \beta_{d\mathrm{M,forum}} - \beta_{d\mathrm{F,forum}}),$$

where each β coefficient has a subscript to denote whether it is the coefficient for the survey data or forum data. Thus, each ratio is increased or decreased by the ratio of the drug-independent demographic ratios between the two data sources. This ensures that the ratios for emerging drugs are normalized relative to the demographic associations for established drugs, so that they can be interpreted more naturally.

Metric	Int.	Female	Male	18-25	26-34	35-49	50+
	Overall						
Survey: Past month	0.767	-0.175	0.168	0.799	0.363	-0.195	-0.960
Survey: >100 days	-0.027	-0.214	0.160	0.625	0.262	-0.161	-0.672
Forum: Subforum	2.937	0.006	0.084	0.164	-0.092	-0.123	-0.038
Forum: Keywords	3.253	-0.030	0.078	0.137	-0.007	-0.079	-0.099
	Cannabis						
Survey: Past month	1.565	**0.080**	**0.373**	**0.456**	0.276	0.211	0.169
Survey: >100 days	1.885	**0.061**	**0.500**	**0.738**	**0.503**	0.191	-0.109
Forum: Subforum	-0.175	**-0.228**	0.130	**0.194**	**-0.248**	-0.091	0.068
Forum: Keywords	0.130	-0.120	**0.108**	**0.158**	-0.075	-0.069	**0.129**
	Cocaine						
Survey: Past month	-0.312	**-0.385**	0.036	-0.156	0.085	0.230	-0.123
Survey: >100 days	-0.322	-0.273	0.006	-0.381	-0.095	0.404	0.017
Forum: Subforum	-0.307	-0.003	**-0.073**	**-0.173**	-0.047	0.144	**-0.154**
Forum: Keywords	-0.106	0.012	-0.078	-0.101	-0.025	**0.133**	-0.047
	Hallucinogens						
Survey: Past month	-0.899	-0.050	**0.200**	0.614	0.032	-0.772	-0.923
Survey: >100 days	-1.373	-0.177	0.037	0.208	-0.470	-0.741	-0.230
Forum: Subforum	-0.153	**-0.087**	0.161	0.220	**-0.081**	**-0.508**	**0.141**
Forum: Keywords	-0.084	**-0.117**	0.133	0.131	-0.072	**-0.192**	**0.034**
	Pain Relievers						
Survey: Past month	0.553	**0.089**	0.045	**0.040**	0.098	**0.192**	**0.089**
Survey: >100 days	0.529	**0.073**	0.080	0.191	**0.192**	**0.087**	-0.095
Forum: Subforum	0.423	0.039	-0.068	**-0.122**	0.150	**0.246**	**0.177**
Forum: Keywords	0.280	**0.044**	-0.041	-0.078	0.083	**0.143**	0.128
	Stimulants						
Survey: Past month	-0.745	0.010	-0.224	-0.003	-0.049	-0.094	-0.384
Survey: >100 days	-0.468	0.016	-0.176	-0.097	0.051	0.027	-0.288
Forum: Subforum	0.404	**0.322**	-0.011	**0.028**	0.169	**0.146**	**-0.250**
Forum: Keywords	0.217	**0.191**	**-0.014**	0.019	0.020	**0.112**	**-0.111**
	Tranquilizers						
Survey: Past month	-0.162	0.081	-0.262	-0.152	-0.079	0.038	0.212
Survey: >100 days	-0.251	0.086	-0.288	-0.034	0.081	-0.129	0.034
Forum: Subforum	-0.193	-0.038	-0.056	0.016	-0.035	-0.060	-0.019
Forum: Keywords	-0.436	**-0.040**	**-0.030**	**0.010**	0.062	**-0.205**	-0.233

Table S4: The coefficients of the demographic association models in equations (1) and (2). The first column ("Int") is the drug-specific intercept β_d independent of demographic group, while the other values are the drug-demographic interaction coefficients for each group, β_{dg} and β_{da}. Positive values indicate an increased likelihood of drug use for that demographic group. The first four rows are the model coefficients independent of any drug, the overall intercept β_0 and the demographic-specific intercepts β_g and β_a. Bolded associations are significant with $p<.05$.

Metric	Int.	Female	Male	18-25	26-34	35-49	50+
	Overall						
Forum: Subforum	1.839	-0.370	0.209	0.131	0.015	0.010	0.005
Forum: Keywords	1.774	-0.370	0.275	0.247	0.073	0.056	-0.281
	Synthetic cathinones						
Forum: Subforum	0.050	0.065	-0.015	**-0.215**	**0.170**	**0.054**	**-0.009**
Forum: Keywords	-0.013	**0.049**	-0.011	**-0.164**	0.208	0.029	**-0.124**
	Synthetic cannabinoids						
Forum: Subforum	0.528	-0.356	**0.098**	-0.002	0.199	**0.274**	**0.315**
Forum: Keywords	0.490	-0.310	**0.088**	-0.005	0.168	0.095	**0.454**
	Phenethylamines						
Forum: Subforum	-0.229	-0.168	0.070	**0.117**	**-0.145**	-0.090	-0.012
Forum: Keywords	-0.385	-0.167	0.190	0.253	**-0.099**	0.004	-0.565
	Salvia divinorum						
Forum: Subforum	-0.348	**0.090**	0.056	0.230	**-0.208**	-0.227	-0.289
Forum: Keywords	-0.092	0.058	0.009	0.162	**-0.204**	-0.071	-0.045

Table S5: The coefficients of the demographic association models in equations (1) and (2) for four emerging drug subcategories, for the two forum metrics. Bolded associations are significant with $p<.05$.

Temporal Associations

Methods

The same linear model in equations (1) and (2) was used for estimating temporal associations with each drug and each year. The prevalence for a given time value t (2007–2012) was modeled as:

(5) $\log y_{dt} = \beta_0 + \beta_d + \beta_t + \beta_{dt}$

Analogous to the demographic model, the β_d and β_t intercepts account for overall prevalence of the drug d and overall activity during time t, while the interaction variables β_{dt} account for associations with specific drugs and specific time periods.

Results: Established Drugs

Table S6 shows the model coefficients for drug associations with each year, again with overall intercepts (β_0 and β_t) and drug-specific terms (β_d and β_{dt}).

The temporal ratios in Table 3 are computed analogously to the demographic ratios, using the method described above in equation (3).

Results: Emerging Drugs

Table S7 shows the temporal regression model coefficients for the four emerging drugs.

Metric	Int.	2007	2008	2009	2010	2011	2012
	Overall						
Survey: Past month	0.151	0.007	-0.015	0.045	0.025	-0.068	0.006
Survey: >100 days	-0.556	0.017	-0.001	0.021	-0.007	-0.063	0.033
Forum: Subforum	3.054	0.269	0.256	0.092	-0.116	-0.235	-0.266
Forum: Keywords	3.322	0.148	0.133	0.089	-0.067	-0.126	-0.178
	Cannabis						
Survey: Past month	1.504	0.113	0.184	0.216	0.279	0.373	0.339
Survey: >100 days	1.751	0.097	0.217	0.242	0.360	0.416	0.420
Forum: Subforum	-0.032	0.096	**0.237**	**0.195**	**-0.068**	**-0.221**	-0.270
Forum: Keywords	0.161	-0.010	**0.043**	0.103	0.063	0.006	**-0.044**
	Cocaine						
Survey: Past month	-0.370	0.213	0.129	-0.182	-0.162	-0.223	-0.144
Survey: >100 days	-0.359	0.205	-0.001	-0.022	-0.282	-0.226	-0.034
Forum: Subforum	-0.295	**0.258**	**0.169**	**-0.072**	-0.123	-0.271	-0.257
Forum: Keywords	-0.152	0.138	0.058	-0.003	-0.087	-0.121	**-0.138**
	Hallucinogens						
Survey: Past month	-0.672	-0.179	-0.158	-0.035	-0.015	-0.104	-0.179
Survey: >100 days	-0.903	-0.167	-0.150	-0.172	-0.143	-0.087	-0.184
Forum: Subforum	0.022	**0.264**	0.145	-0.017	0.010	-0.153	-0.228
Forum: Keywords	0.050	0.188	0.134	0.023	0.003	-0.136	-0.162
	Pain Relievers						
Survey: Past month	0.490	0.140	0.066	0.102	0.122	0.015	0.045
Survey: >100 days	0.416	0.018	0.035	0.118	0.147	0.098	0.001
Forum: Subforum	0.265	-0.176	-0.059	0.004	0.066	0.214	0.216
Forum: Keywords	0.212	-0.037	0.033	-0.000	0.011	0.111	0.094
	Stimulants						
Survey: Past month	-0.698	**-0.153**	-0.132	-0.009	-0.172	-0.078	-0.153
Survey: >100 days	-0.555	**-0.109**	-0.092	-0.114	-0.085	-0.029	-0.126
Forum: Subforum	0.276	**-0.164**	**-0.186**	**-0.018**	0.046	**0.221**	0.377
Forum: Keywords	0.152	-0.010	-0.072	0.015	-0.005	0.094	**0.130**
	Tranquilizers						
Survey: Past month	-0.255	-0.126	-0.104	-0.046	-0.026	-0.051	0.098
Survey: >100 days	-0.350	**-0.027**	-0.010	-0.032	-0.003	-0.235	-0.044
Forum: Subforum	-0.235	**-0.009**	**-0.050**	**-0.000**	**-0.047**	**-0.025**	-0.103
Forum: Keywords	-0.423	-0.121	**-0.063**	-0.050	**-0.052**	-0.079	**-0.058**

Table S6: The coefficients of the temporal association model in equation (5). The first column ("Int") is the drug-specific intercept β_d independent of temporal group, while the other values are the drug-year interaction coefficients for each year, β_t. Positive values indicate an increased likelihood of drug use in that year. The first four rows are the model coefficients independent of any drug, the overall intercept β_0 and the temporal-specific intercepts β_t. Bolded associations are significant with $p<.05$.

Metric	Int.	2007	2008	2009	2010	2011	2012
	Overall						
Forum: Subforum	1.810	0.101	-0.118	0.112	0.133	0.006	-0.235
Forum: Keywords	1.844	0.097	-0.072	0.134	0.156	0.005	-0.319
	Synthetic cathinones						
Forum: Subforum	-0.120	**-0.443**	**-0.655**	**0.186**	**0.339**	**0.294**	**0.159**
Forum: Keywords	-0.289	**-0.696**	**-0.728**	**0.005**	**0.314**	**0.498**	**0.317**
	Synthetic cannabinoids						
Forum: Subforum	0.216	**-0.827**	**-0.582**	-0.123	0.613	0.606	0.529
Forum: Keywords	0.261	**-0.608**	**-0.233**	-0.068	0.455	0.397	**0.320**
	Phenethylamines						
Forum: Subforum	-0.013	**0.450**	**0.310**	**-0.155**	**-0.440**	**-0.238**	**0.059**
Forum: Keywords	-0.060	**0.624**	**0.243**	**-0.049**	**-0.276**	**-0.244**	**-0.358**
	Salvia divinorum						
Forum: Subforum	-0.083	**0.921**	**0.809**	**0.204**	**-0.379**	-0.655	-0.983
Forum: Keywords	0.088	**0.777**	**0.646**	**0.246**	**-0.336**	**-0.647**	**-0.597**

Table S7: The coefficients of the temporal association model in equation (3) for four emerging drugs, for the two forum metrics. Bolded associations are significant with $p<.05$.

Demographic and Temporal Word Associations

Methods

As an additional experiment, associations between the words used in forum messages and the demographic/temporal groups are modeled. This is a way to examine, at a high level, trends in the *content* being written by various groups during various periods.

Let y_{wi} be the number of times the word w was posted in the forum within group i (either a gender, age group, or time period).

The log of the word frequencies for each gender (g) and each age (a) are modeled as:

(6) $\log y_{wg} = \beta_0 + \beta_w + \beta_g + \beta_{wg}$
(7) $\log y_{wa} = \beta_0 + \beta_w + \beta_a + \beta_{wa}$

Similarly, word frequencies for each year t are modeled as:

(8) $\log y_{wt} = \beta_0 + \beta_w + \beta_t + \beta_{wt}$

The β_w variables are word-specific intercepts which capture a word's overall likelihood of being written, independent of the demographic/temporal group. As with the drug association models, β_g, β_a and β_t are group-specific intercepts, which in this case adjust for differences in overall word frequencies among these groups, since some groups have fewer messages. The β_{wg}, β_{wa} and β_{wt} variables are interaction terms which capture associations between specific words and specific groups. These models are not restricted to specific drugs and are estimated on all 45 drug subforums.

The resulting interaction coefficients provide qualitative insight into differences in the content of messages from different demographic groups and time periods.

Results: Demographic Associations

Table S8 shows the 50 words with the highest coefficients for each demographic group. The word associations are estimated among thousands of words used in the forums, so the coefficients are inevitably "noisy" and not all word associations have meaning. Nevertheless, these results offer a way to understand key associations between groups that are inferred from tens of thousands of forum messages.

Many of the terms refer to specific drugs, and the word associations largely match the associations with drug categories presented in the previous sections. In some cases, specific word associations reveal finer grained distinctions between drugs than the broader categories analyzed above. For example, among drugs included in the Stimulants category, the terms "adderall", "ritalin" and "stimulants" have stronger female associations, while the terms "amphetamine(s)", "ecstasy" and "mdma" have stronger male associations. The terms "adderall" and "amphetamine(s)" have stronger associations with the 18-25 years age group, while "methamphetamine" has a stronger association with the 35-49 years age group.

The top words also reveal associations with drugs that were not analyzed in the study. For example, terms related to dextromethorphan ("dxm", "syrup") are heavily associated with the 18-25 years age group.

A general difference between genders is that the male word associations include a much larger number of different drugs, including references to emerging drugs like 2C-phenethylamines and *Salvia divinorum*. The specific drugs included in the female word associations are predominantly stimulants and opioids.

Other word associations simply reflect characteristics of the demographic group independent of drug use. For example, forum members are not allowed to incriminate themselves and often write in third person, so it makes sense that gendered pronouns such as "shes" and "hes" (apostrophes removed) are heavily associated with the respective genders.

Results: Temporal Associations

Table S9 shows the 50 words with the highest coefficients for each year.

An interesting characteristic of the year-specific associations is that many of the top associations are references to the four emerging drugs this study analyzed. At least one term related to an emerging drug is among the top five words for each year except 2012. That these terms have strong temporal associations, even though this model used data from all drug subforums, shows that these drugs in particular have high temporal variability, in agreement with the study's results from the previous sections.

Terms related to synthetic cannabinoids ("jwh-018", "blends") have very high associations with 2009 – 2011, as do terms related to synthetic canthinones ("mephedrone", "mdpv"). Terms related to *Salvia divinorum* ("salvia") are strongly associated with 2007 – 2008, and terms related to phenethylamines ("2c-e", "2c-i") are strongly associated with 2007. These associations align with the trends from the drug association models.

Female	Male	18-25	26-34	35-49	50+
shes	2c-e	dxm	alprazolam	2c-e	hash
haha	dxm	haha	ghb	mephedrone	patch
opioids	gbl	xr	kratom	ecstasy	acetone
heroin	ghb	alprazolam	dr	2c-i	doc
methadone	alprazolam	plateau	methylphenidate	acetone	pain
veins	2c-i	adderall	dxm	crystals	dude
ritalin	xr	2c-e	clonazepam	mdpv	ethanol
adderall	salvia	ghb	anyways	mdma	girl
theyre	haha	amphetamines	blends	research	oil
youre	dmt	2c-i	really	compound	ice
xr	effects	trip	afoaf	lab	years
opioid	trip	xanax	benzo	methylone	wash
adhd	plateau	heroin	plateau	chemicals	hi
really	adderall	amphetamine	people	drugdrugs	peace
vein	hes	clonazepam	suboxone	duration	bags
stimulants	amphetamines	gbl	2mg	reports	pods
patient	amphetamine	hes	going	news	opiates
people	methylphenidate	syrup	know	ghb	dry
needle	ecstasy	tripping	time	ml	tar
disorder	dopamine	effects	day	info	seeds
fucking	8217	dopamine	beer	cannabinoids	25
know	serotonin	benzos	blend	rcs	edited
time	duration	felt	alot	jwh-018	dried
medications	mdma	high	tea	legal	mg
theres	nitrous	really	effects	00	relief
ssris	benzodiazepines	friend	benzos	law	grams
helpful	experience	ur	stuff	synthetic	alkaloids
oh	lsd	serotonin	make	testing	chronic
person	cannabis	weed	pretty	product	old
syringe	psychedelic	benzo	terrible	rc	pm
id	receptor	time	way	dr	filter
going	dose	friends	probably	patch	hcl
feel	monkey	1mg	sure	drug	solution
okay	opium	school	drink	threads	paper
sick	visuals	feeling	feel	cocaine	expensive
sort	time	visuals	say	act	cannabis
dope	high	wondering	person	methamphetamine	marijuana
meth	clonazepam	feel	sugar	results	fresh
ones	drug	ecstasy	need	monkey	oxy
drugs	syrup	experience	drugs	paper	buprenorphine
thats	trips	trips	want	number	thc
think	mescaline	drug	ritalin	search	dat
make	weed	fucked	safe	unknown	tabs
dealer	alot	amazing	hours	wash	dissolve
want	1mg	feels	think	uk	meds

Table S8: The 45 highest word associations for each demographic group. All words have been lowercased and punctuation has been removed.

2007	2008	2009	2010	2011	2012
nitrous	en	jwh-018	jwh-018	afoaf	afoaf
en	dmt	blends	mephedrone	jwh-018	ive
ml	8217	magic	swimmers	dog	ur
salvia	alprazolam	swimmers	afoaf	cat	meth
van	salvia	mephedrone	pods	mdpv	dog
crystals	thc	gbl	dope	pet	seconds
marijuana	erowid	opium	shes	blends	youre
2c-e	alkaloids	pods	blends	ive	know
data	opioids	spice	op	blend	havent
studies	acetone	tar	heroin	heroin	think
cannabis	pot	thinks	thinks	friend	ill
study	material	heroin	time	know	feel
erowid	extraction	mescaline	dose	seconds	want
mescaline	methylphenidate	poppy	hes	day	later
morphine	bong	agrees	high	cannabinoids	day
hcl	extract	dope	feels	id	synthetic
ketamine	psychedelic	tea	knows	pain	minutes
base	effects	time	dog	youre	added
opium	drugs-forum	hes	methadone	withdrawals	pain
ghb	visual	mdma	dxm	mephedrone	adderall
cocaine	opioid	knows	first	methadone	need
mixture	forum	vein	really	feel	days
administration	crystals	finds	drug	think	friend
tramadol	lsd	shes	effects	synthetic	id
dxm	mdma	really	pet	doctor	time
reported	active	pills	anxiety	later	going
plant	experience	effects	patient	op	really
compound	tramadol	amphetamine	friend	addicts	help
driving	ml	shot	people	withdrawal	lol
0	dr	high	experience	days	youve
acetone	clonazepam	feels	using	going	stop
effects	edit	way	gear	want	thank
produced	mescaline	needle	wants	time	hope
2c-i	tobacco	day	way	said	started
alcohol	likely	likes	day	kratom	doing
dopamine	nitrous	bag	finds	life	guess
growing	rules	blend	know	shes	life
duration	cocaine	best	going	added	guys
article	present	prefers	night	help	high
comments	water	dose	feel	minutes	kratom
extraction	trip	probably	needle	really	wanna
leaves	experiences	people	opiates	high	wish
alot	00	first	friends	mg	honestly
include	plants	veins	phone	quit	people
lsd	time	bit	hed	people	hours

Table S9: The 45 highest word associations for each year of data. All words have been lowercased and punctuation has been removed.