Twitter Feeds and Google Search Query Surveillance:

Can They Supplement Survey Data Collection?

Joe Murphy, Annice Kim, Heather Hagood, Ashley Richards, Cynthia Augustine, Larry Kroutil, and Adam Sage – RTI International

Abstract

With networks like Twitter allowing for mass-scale sharing of thoughts, opinions, and behaviors by people worldwide, there is an opportunity to harvest these data to provide insights where survey data have traditionally been employed. Similarly, trends in Internet searches can be tracked over time, and studies have shown some correlation of search trends with results of political polls, flu outbreaks, and other phenomena. It is worth considering whether and how new sources of data may supplement traditional survey data collection. In this paper, we consider these data sources as they relate to the current and future states of survey data collection. We explore the surveillance of an emerging drug trend for *salvia divinorum,* using Twitter feeds and Google search trends, and compare tweet content and Google search volume to reports from publicly available survey data. We highlight the results of these comparisons and discuss the implications for future research in the survey context.

1. Introduction

With the growing prevalence of online social networks, sources of data are now available that were barely imaginable 10 years ago. Networks like Twitter allow for mass-scale sharing of thoughts, opinions, and behaviors by people worldwide. There is an opportunity to harvest these data to provide insights where survey data have traditionally been employed. In the United States, 79% of adults go online and 61% visit social media sites (Pew Internet & American Life, 2011). Among teens aged 12 to 17, 93% are online and 74% have a profile on a social networking sites, and social networking is the online activity on which they spend the most time daily (Pew Internet & American Life, 2011; Kaiser Family Foundation, 2010). Simultaneously, survey researchers are facing difficulties in efficiently collecting data because of decreases in landline telephone coverage and willingness to participate by respondents. The attractiveness

of using available data from passive online sources is enhanced by the relative costliness and time-intensive nature of survey research.

In recent years, researchers have begun analyzing massive volumes of data from Twitter and search queries from Google to demonstrate that people's information-seeking and information-sharing behaviors online are correlated with phenomena such as flu outbreaks, introduction of new nicotine delivery products, tobacco tax avoidance, and even weekend box office sales for movies (e.g., Ginsberg et al., 2009; Ayers et al, 2011a, 2011b; Goel et al., 2010). These studies exemplify the emerging fields of *infoveillance* and *sentiment analysis*. Infoveillance is the automated and continuous analysis of unstructured, free text information available on the Internet (Eysenbach, 2009). Sentiment analysis is the automated computational coding of text to determine if expressed opinions are positive, neutral, or negative. Applied to sources like Twitter postings (tweets) and Google searches, these approaches have been shown, in some circumstances, to produce results that correlate with and even predict those collected through traditional data collection methods. For example, Eysenbach (2009) found a high correlation between clicks on sponsored search results for flu-related keywords and actual flu cases using epidemiological data from 2004 to 2005 and Polgreen and colleagues (2008) showed that search volume for influenza-related queries was correlated with actual reported flu from 2004 to 2008.

As a supplement or alternative to traditional survey research, infoveillance has some attractive qualities. The sheer volume of Google search queries and postings on Twitter provide a glimpse into the thoughts of at least some subset of the general population. As evidenced by prior studies, these streams have the potential to replicate health trends, often providing an earlier indication of trends than what can reasonably be supplied via surveys. With data available on the Web, no burden is placed on survey respondents because there are no respondents, as traditionally defined.

The speed at which one can investigate a topic of interest using infoveillance greatly exceeds that of a traditional survey approach. Whereas a survey requires sample identification, question construction, contact attempts, and data collection prior to analysis, infoveillance requires only access to the stream of queries and postings and a methods for analyzing the content.

Infoveillance suffers, however, from a high degree of obscurity around several of the most important tenets of survey methodology. The degree to which queries and postings represent the general population, let alone a specific target population for any given study, is practically unknown. The lack of information on the coverage of these data makes it impossible to

construct accurate population-based estimates. Those who search or post are not guided in any way that is related to the information sought. In essence, the respondent is making up his or her own questions to a "survey," and thus, there is no standardization or check on the validity of the information being shared.

How do the results of these studies compare to estimates obtained though traditional survey methods? What does it mean if these new methods can produce results that correlate with, or even appear to predict, those collected through traditional survey methods, but the representativeness and other error properties of tweets and search queries are unknown?

Take, for example, an important survey topic such as substance abuse. The traditional approach employed by studies such as the U.S.-based National Survey on Drug Use and Health (NSDUH) involves the listing of housing units across the country, precise sampling of units and respondents, in-person visits to almost 180,000 households by trained field interviewers, administration of a carefully constructed and tested 1-hour questionnaire, compilation and cleaning of responses, application of survey weights, and analysis to produce annual official estimates of the use and abuse of both legal and illicit drugs, among many other related topics (SAMHSA, 2010a, 2010b). The survey aims, in part, to identify emerging drugs of abuse that may be gaining popularity, which can be important in understanding trends and the need for public education and treatment.

An infoveillance approach to the issue of emerging drugs would differ greatly. Armed with basic information on a particular substance of interest, a researcher could mine the stream of queries and postings to detect the level at which people have been and are posting information or thoughts about this substance on the Internet. Although the motivation for sharing these thoughts is not directly evident, one can measure the level of Internet activity as it varies over time and the sentiment behind the shared information (i.e., whether it portrays the substance in a positive, neutral, or negative light).

In this paper, we consider the above questions as they relate to the current and future states of survey data collection. We explore one potential application in particular—the surveillance of an emerging trend in use of *salvia divinorum* by using Twitter feeds and Google search data. *Salvia divinorum* is an herb common to Central and South America. Recreational users typically smoke the dried leaves. The active ingredient is salvinorum A, which has hallucinogenic properties. The drug is not currently regulated under the Controlled Substances Act, but several States have passed legislation to regulate its use. *Salvia divinorum* received considerable media attention toward the end of 2010 when a video of the popular singer Miley Cyrus smoking

the drug was posted on YouTube. For the remainder of this paper, we refer to *salvia divinorum* simply as salvia.

We are particularly interested in emerging drugs—those with a low rate of use that may increase in popularity—because traditional survey methods require a very large sample to detect their emergence and often are not able to report the trend until a rate has spiked. Infoveillance has the potential to detect trends for rare but emerging drugs and it does it so much quicker than a traditional survey approach.

Specifically, we address these questions:

1) Is information-seeking behavior (Google searches) about salvia associated with actual self-reported use of these drugs?
2) Is information-sharing behavior (tweets) about salvia associated with actual self-reported use of these drugs?
3) What are the main topics and sentiment of tweets for salvia? How do automated text analytics tools compare with manually coding unstructured Twitter data?

We compare Google search volume and content of tweets for salvia use with reports from publicly available survey data. We highlight the results of these comparisons and discuss the strengths and limitations of infoveillance as applied to survey research and the implications for future research.

2. Data Sources

Twitter

Twitter is a free online social networking and microblogging service where users can send and read tweets or short messages up to 140 characters. Twitter has over 300 million registered users (tweeters) and more than 140 million tweets per day (Rowinski, 2011; Twitter Blog, 2011). A recent Pew study of Internet users reports that 18% of 18 to 24 year olds and 19% of 25 to 34 year olds use Twitter (Smith, 2011). Overall, 13% of online adults use Twitter and half of them access it using their mobile phone (Smith, 2011).

We conducted a review of available social media monitoring tools for use in this study. We compared four potential vendors on price, scope of Twitter access, and availability of historical data. Of the tools reviewed, Radian6 had the most comprehensive coverage of Twitter data, which includes full access to all tweets from May 1, 2008 to present.

Relevant tweets about salvia were identified by Radian6 using “salvia” as a keyword search. The Radian6 conversation cloud, which is a visual representation of the most common words that appear in tweets retrieved, was used to refine the keywords. For example, a portion of tweets were about gardening, as varieties of salvia other than *salvia divinorum* are commonly found in the garden, so we excluded references to “garden” and “gardening.” We also restricted the Radian6 twitter data to the English language for coding and to the United States for comparability with the drug use survey data.

Google Insights for Search and News Archive

Google Insights for Search is a free tool provided by Google that can be used to monitor trends in public keyword search queries. Google Insights provides the “likelihood” that a Google search engine user will search for a term over a given time period or geographical location. Search term queries can be filtered by search type (i.e., Web, image, news, or product), geography (i.e., country, subregion, or metro), time range, and category (e.g., automotive, health, real estate, travel). Google Insights data are presented as a relative scale. To compute the scale, Google uses a proportion of de-identified search data. Frequencies of each search term are then divided by the total number of searches. The data are then normalized by common variables, such as population, so terms can be compared across geographic locations (Google Insights for Search, 2011a). Once the search data are normalized, each point is divided by 100, producing a relative scale from 0 to 100 (Google Insights for Search, 2011b). Historical data were available from Google Insights starting January 1, 2004.

To examine the volume of searches around salvia, we used the keyword “salvia,” but excluded references to garden(ing) and limited the search to queries from the United States. We downloaded Google Insights data from January 1, 2004 to June 30, 2011 into comma separated (.csv) files for analysis.

To better understand what events may have influenced people to search for salvia information, we extracted headline news using Google News (2011), a free tool that crawls and identifies relevant content from their online database of current and historical newspapers, news sources (i.e., radio and television news), magazines, and legal documents.

We searched for relevant news when the volume of salvia searches spiked (Google Insights=100). This was done by using the keyword “salvia” in Google News with custom date ranges spanning a week (from Saturday to Sunday) around a spike in Google Insights.

Searches were restricted to all available English language headlines, and we only examined free content retrieved.

National Survey on Drug Use and Health (NSDUH)

The NSDUH is an annual survey sponsored by the Substance Abuse and Mental Health Services Administration (SAMHSA), an agency of the U.S. Department of Health and Human Services. The survey provides representative estimates of the use of tobacco, alcohol, illicit drugs (including nonmedical use of prescription drugs), and mental health in the United States at the national, State, and substate levels.

The NSDUH public use files from 2004 through 2009 were used in this analysis. The data consist of core and noncore (i.e., supplemental) sections. The core set of questions critical for basic trend measurement of prevalence estimates remains in the survey every year and comprises the first part of the interview. Although specific questionnaire items about salvia were included in a noncore section since 2006, reports of salvia use can be identified from the "other" response options for core drug questions. For example, the survey contains a core section about hallucinogenic drugs that does not explicitly ask about salvia. At the end of the section on hallucinogens, the respondent is asked "Have you ever, even once, used any other hallucinogens besides the ones that have been listed?" If so, the respondent may use a free-response section to report use of up to five additional drugs. The written responses are subsequently assigned numeric codes corresponding to specific drugs.

3. Methods

Compilation and Analysis of Twitter & Google Data

Weekly Google Insights search data were assembled for the time period of January 1, 2004 to June 30, 2011. For comparison, the proportion of tweeters was calculated by taking the total monthly number of tweeters mentioning salvia at least once and dividing by the estimated monthly number of registered tweeters. The monthly number of registered tweeters was modeled using the number of registered users identified in the twitter blog (http://blog.twitter.com/). An exponential model was fit to estimate the remaining monthly registered tweeter numbers. The proportion of tweeters was used instead of tweet volume to account for the growth in the population of tweeters and to represent the data at the person

level, similar to NSDUH data on percent use. Twitter data, Google Insights search data, and Google News headlines were then compiled using Excel. A linear trend line was fitted to the Twitter and Google Insights search data to assess whether tweets and searches for saliva have increased significantly over time.

Text Analysis of Twitter Data

We used IBM SPSS Text Analytics for Surveys (STAS) to code tweets. STAS is a software program designed to categorize text responses in surveys. This allows for open-ended text to be quantified for analysis with other survey data. The process is automated, but results are fine-tuned by the user who trains the software to code text in the particular data set. The benefits of using this software include improved efficiency and consistency when coding text.

This software is also useful for coding tweets because tweets are brief, much like open-ended responses to survey questions. In addition, because STAS is a highly trainable software, this was especially useful for coding twitter data, which are challenging to work with because slang and emoticons are more common than in other forms of written communication.

STAS uses an extraction method to automatically identify common words or phrases in the data. The prevalent terms are then assigned to categories for coding. In this analysis, each code was treated as a dichotomous variable. Because of the high number of tweets in the data set, the analysis was conducted for only a random sample of 500 tweets.

Sentiment Analysis of Twitter Data

The next step in the coding process was to code the tweets according to sentiment (positive, neutral, or negative). Both STAS and Radian6 provide automated sentiment analysis of tweets, which largely involves looking for the use of certain words that have been categorized as expressing either positive, neutral, or negative sentiment. Because of the complexity of how language is used in tweets (e.g., sarcasm, hyperbole, emoticons), we were skeptical of the automated analysis approach, so we manually coded a random sample of 500 salvia tweets to compare the manual approach (gold standard, but resource intensive) to STAS and Radian6's automated coding (time efficient, but unknown accuracy). We created a codebook to define what tweets would constitute positive, neutral, and negative sentiment. Tweets were assigned the positive code if they endorsed the use of salvia, justified that the drug is safe to use, or

referred to the drug positively in other ways. Tweets were also coded positive if the tweeter used the drug and did not have a bad experience or expressed interest in buying or selling the drug. Tweets were coded negative if the person tweeting expressed opposition or disapproval to the use of salvia, or the tweeter used the drug and expressed negative feelings about it.

The neutral code was used for most of the remaining tweets. Tweets were coded neutral if they mentioned salvia, but the drug was not the main topic of the tweet, or if the tweeter referred to the drug without expressing his or her own opinion about it (e.g., sharing a news story about salvia).

Tweets were coded irrelevant if they did not refer to the drug salvia. This was typically due to misspellings (e.g., 'salvia' misspelled 'saliva') or alternate uses of the words (e.g., Salvia Street as part of an address). Tweets were labeled not codable if they contained only symbols or a link. Two independent coders assigned codes and an adjudicator made final decisions when the coders disagreed.

We compared the manually coded sentiment, which we considered the gold standard, with the automated sentiment from Radian6 and STAS. This comparison was made for 466 salvia tweets. These were the tweets remaining after dropping irrelevant and uncodable tweets from the initial samples of 500 that were manually coded.

The Radian6 sentiment was preassigned to the tweets we downloaded. For STAS sentiment we used the built-in libraries of 5,677 positive and 4,834 negative terms that would have been the foundation of our coding system had we decided to set up a coding system in STAS. Tweets including both positive and negative terms were coded mixed and tweets with none of these terms were coded neutral.

NSDUH Survey Analysis

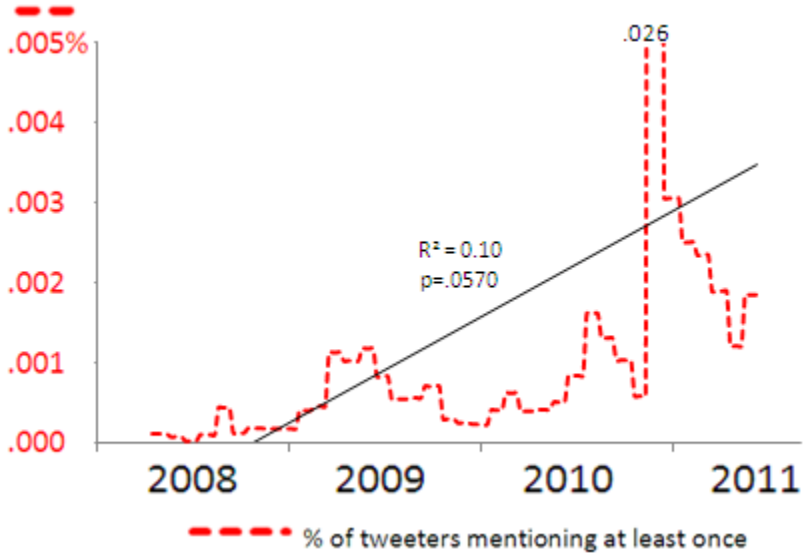
Although NSDUH is a national survey that can be analyzed with statistical weights to represent the national population, there were very few respondents reporting use of salvia. For our analysis, we computed unweighted counts of salvia use by quarter. Although the NSDUH is conducted throughout the year, the public use data file only provides quarterly timeframes. To protect respondent privacy, there is very little demographic information in the public use files so no stratified analysis (e.g. by respondents' geographic location) could be conducted.

4. Results

Appendix A provides an overview of the time trends in salvia searches, tweets, and use, along with salvia-related headline news. We summarize results for each data source in more detail below.

As shown in Figure 1, the proportion of tweeters sharing information about salvia on Twitter increased from 2008 to 2011 (nearly significant: p-value=.0570). This was largely driven by the huge spike in activity around December 2010 when videos of celebrity Miley Cyrus smoking salvia circulated online.

Figure 1. Proportion of Tweeters Sharing Information each Month, 2008 to 2011



Regarding Google searches, Figure 2 shows there is a slight positive trend in the likelihood of users searching for saliva from January 1, 2004 to June 30, 2011. The peaks in search index during 2008 correspond to news stories about the increasing popularity of salvia, the health effects of salvia use, and lawmakers considering regulation of salvia. By far, the largest surge in search activity occurred around December 2010 when videos of Miley Cyrus smoking salvia circulated online.

Figure 2. Salvia Google Search Volume by Week, 2004 to 2011

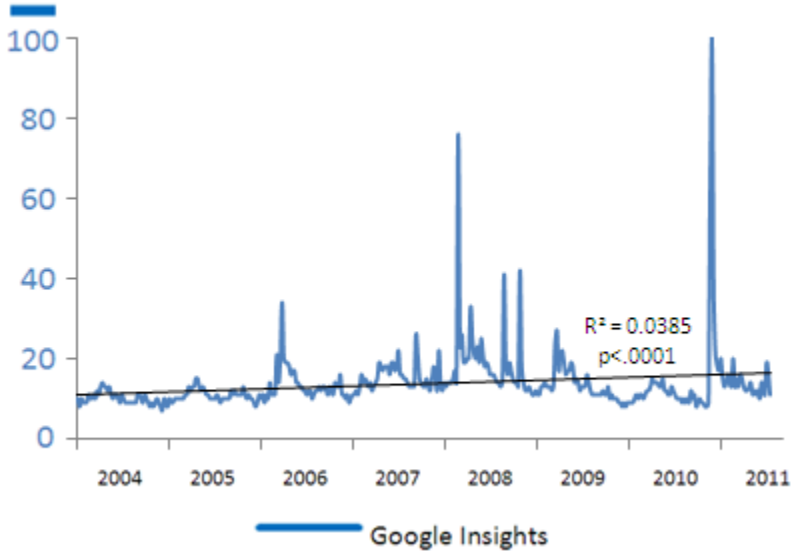
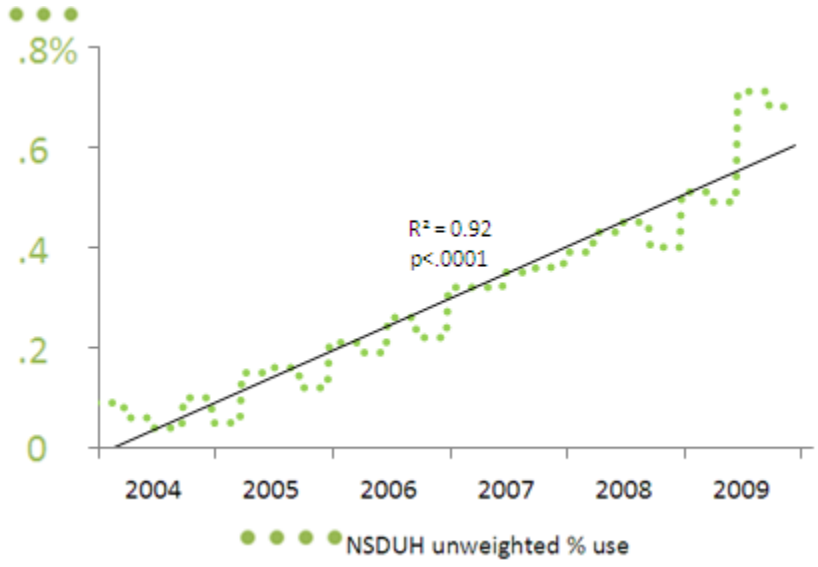


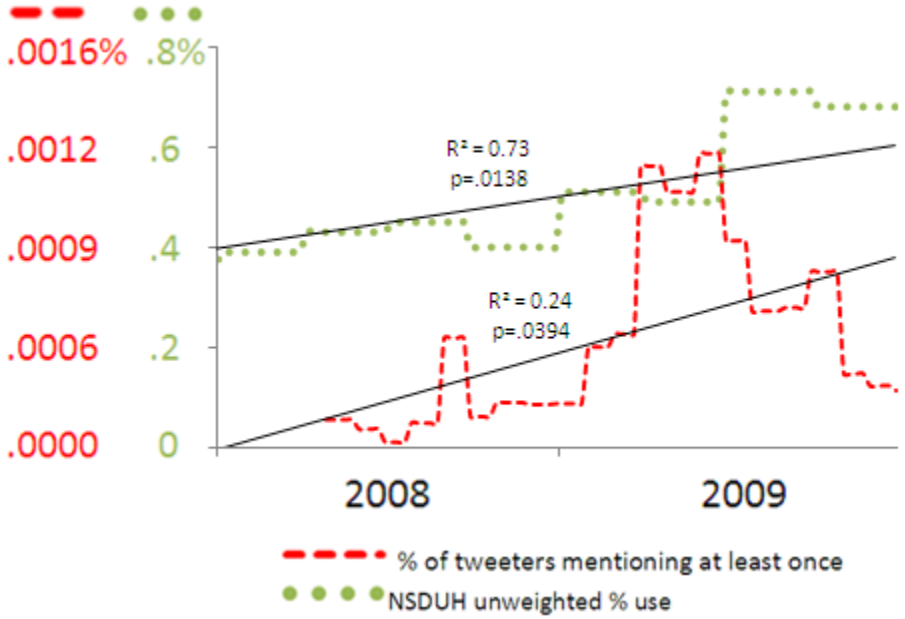
Figure 3 shows that based on NSDUH other-specify data, the lifetime prevalence of salvia use was low between 2004 and 2009, with less than 1% of respondents reporting the drug. However, there has been a slow but steady significant increase in salvia use from 2004 to 2009 (R^2=0.92, p-value < 0.0001).

Figure 3. Salvia Use, Quarterly Estimates, 2004-2009 NSDUH



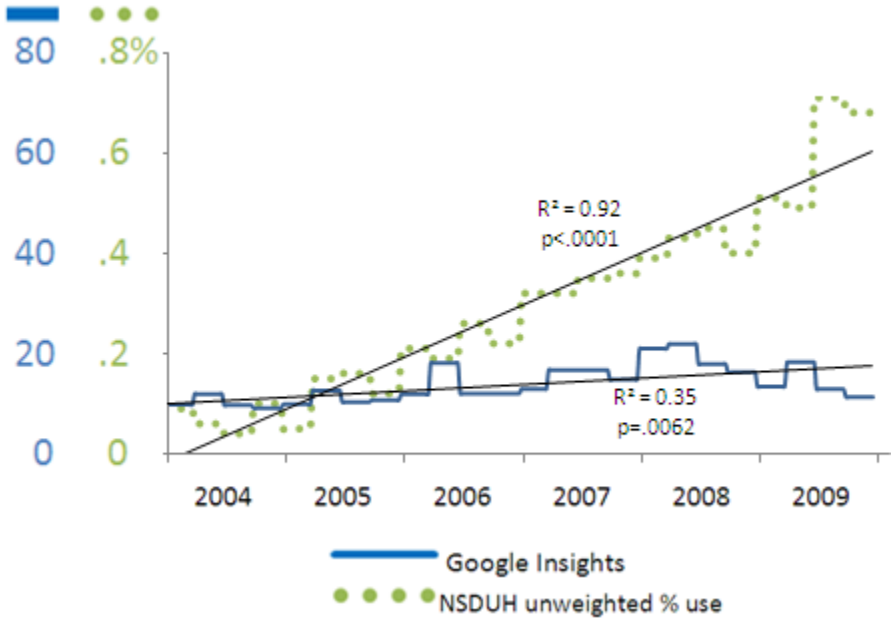
We charted NSDUH use and tweets on a common graph and fit linear trend lines for the period where both sets of data were available (2008 to 2009), revealing similar positive trends. It is interesting to note that an increase in the proportion of tweeters from April to June 2009 is followed by an increase in NSDUH percent use for the third quarter of 2009.

Figure 4. Tweets about Salvia vs. Self-reported Salvia Use (NSDUH), Quarterly, 2008 to 2009



When we examined self-reported salvia use and Google search data, the two trends did not appear to correspond closely with one another. For comparability with NSDUH, we show quarterly Google data in Figure 5. The significant (p-value <.0001) increase in mentions of salvia in the NSDUH correspond with only a slight increase in Google search volume over the period of 2004 to 2009.

Figure 5. Salvia Google Searches and Self-reported Salvia Use (NSDUH) by Quarter, 2004 to 2009



Regarding the sentiment of tweets, our manual coding suggests that salvia tweets most commonly concerned use of the drug (55%), Miley Cyrus (47%), links with additional information (39%) (see Table 1). In contrast, few tweets were about the safety and health effects of salvia use (3%) and medical conditions for salvia use (1%). Our manual coding of sentiment found that, among relevant tweets, only 21% portrayed salvia in a positive light, Most of the tweets were neutral (61%), and 16% described salvia in negative terms. When we compared the automated sentiment analysis to our manual coding, the results were not promising. STAS' total match rate with our manual coding was 44% and Radian6's was 59%. This means that about half the time, the automated coding produced a result different than the manual coding, which we assumed to be the gold standard.

Table 1. Top Most Common Topics for Salvia Tweets (N=500)

Topic	% of Salvia Tweets	% Positive	% Neutral	% Negative
Using the drug (e.g., high on salvia)	55	14	86	0
Miley Cyrus	47	12	62	26
URL Included (e.g., link to news article, video, online store)	39	14	82	4
Video Reference (e.g., YouTube)	13	18	78	4
Legal Issues (e.g., legislation, legality)	8	25	72	3
Buying and/or Selling	5	65	35	0
Gardening (e.g., plants, seeds)*	5	19	81	0
Research (e.g., drug safety, harmful effects, health concerns)	3	33	61	6
Uses for the Drug (e.g., depression treatment, horse tranquilizer, migraine cure)	1	23	50	27
All Tweets	100	21	61	17

*The Radian6 search excluded tweets with the words "garden" and "gardening," but other gardening references remained (e.g. *I don't have to water flowers today! No frost yet, but I pulled out my sick salvia and impatiens - too leggy.*)

5. Discussion

In this paper, we investigated the potential for mining Twitter and Google search data to study the emerging trend in salvia drug use. Twitter and Google data both showed increases in the presence of salvia, and while use remained rare over the time period studied, NSDUH data suggest it did increase over time.

We see some evidence that information sharing about salvia on Twitter may be associated with actual self-reported salvia use. Although the trends in tweeting and salvia use follow similar patterns, more formal statistical testing is needed to determine whether there is a significant relationship between these trends. The period of overlap for both sources was rather short and a better picture might arise when 2010 NSDUH data are available, especially given the large spike in tweet volume in December 2010 related to the Miley Cyrus video. In contrast, Google and NSDUH data did not appear to correspond highly.

Our results suggest that people who tweet about salvia are mostly sharing information about salvia use, including the YouTube video in which celebrity Miley Cyrus was shown smoking

salvia and reacting to its effects. The peak in tweets about Miley Cyrus and the corresponding headline news is not surprising. Previous studies have shown that celebrities are influential in driving news coverage about health topics (e.g., Chapman et al., 2005). Given Cyrus' teen idol status, the YouTube video and corresponding conversations on Twitter might influence youth awareness and curiosity about salvia. Cyrus' influence can be examined in better detail once the public use NSDUH data for 2010 are released.

The results of our sentiment analysis suggest that current automated methods do not replicate the gold standard of manual coding. STAS and Radian6, like most vendors and applications that provide automated sentiment analysis, merely scan tweets for the mention of specific words that have been categorized as positive, negative, or neutral. In the manual coding process, we had to iteratively update our codebook as we encountered nuances in tweets, such as sarcasm, that were challenging to assign categorically as positive, negative, or neutral without making some a priori decisions for how to interpret certain phrases or text abbreviations (e.g., 'lol,' 'smh'). Unless vendors and software applications provide options to customize sentiment analysis with more nuanced algorithms, our results suggest that manual coding is still the best approach for analyzing unstructured text. The major drawback to manual coding is the time required to complete the operation. This is especially a concern with the exponential growth in social media sharing and resulting explosion in available data.

Although previous studies have suggested that mining publicly available Twitter and Google search data can help uncover trends in actual behavior, our results suggest that these infoveillance methods may not be universally applicable across topics of study, including drug use. For one, drug use is a rare event and people may not share and seek out information about drug use as they do for other health topics, like flu outbreaks and STDs. Flu and STDs are infectious by nature and transmitted by human contact, which may strongly influence people's willingness to share and seek out information to prevent exposure, assess symptoms, or treat illness. In contrast, drug use does not have the same communal properties of flu outbreaks. Secondly, people may be less willing to discuss or search for information about illegal behaviors, such as drug use, and therefore, the proportion of tweets and volume of Google searches about a particular drug may be lower bound estimates of the population's awareness and interest in a drug. Thirdly, the relatively slow diffusion rate of new drugs of abuse may not be easy to detect with Twitter and Google data. In the case of an infectious disease like influenza, someone can start a chain of infection going through casual contact. Deciding to use a less commonly used (but emerging) substance requires access to a social

network that has access to that substance. Being exposed to information about salvia online may be not be sufficient to influence someone to try salvia if access to the drug is limited.

Although our results showed that conversations about drug use may be driven by key events, such as celebrity mishaps, news of a celebrity being caught for use of a certain drug may not have the same effect on people's interest as a report of a celebrity being diagnosed with a disease. In the latter case, the view might be, "if she/he can get that, then so can I," thereby increasing their perceived susceptibility and interest in obtaining and sharing more information. In contrast, news about celebrity drug use may merely be dismissed by the larger population as a case of another celebrity mishap. In our analysis, more negative than positive sentiments were expressed in tweets about Miley Cyrus. More in-depth analysis of tweets can help illuminate whether celebrity events like Cyrus' serves as teachable moments or problematic social modeling for teens.

There are limitations of the data sources themselves. For example, the demographics of Twitter users do not represent the general population, and only 13% of online adults use Twitter. In addition, little is known about the characteristics of Twitter users who actually tweet about drugs or other health-related topics and at what volume, without actually surveying a representative sample of Twitter users. Therefore, the coverage properties that can be calculated for representative surveys are largely unknown for these new data sources. For these reasons, Twitter data currently do not provide a representative picture of how the population shares information online. However, as the population of Twitter users continues to grow, it is possible that this rich data source may provide a representative pulse of how users share information online.

Compared to a traditional survey approach, data sources like Twitter and Google provide inexpensive option for gaining insight into an emerging topic or putting some context around a point estimate of interest. They are relatively inexpensive and quick to analyze. And with increasing issues related to declining response rates and decreased landline telephone coverage, these social media represent a source of data where people are actively sharing more information. Without interaction between a survey organization and respondent, there is no respondent burden associated with social media analysis, simply because there are no "respondents" in the traditional sense. However, although social media have the potential to supplement survey research, we do not believe, at this point, they can substitute for traditional survey approaches where precise estimates and correlations between standardized measures are needed. Future research could include further analysis of demographic correlates of

substance use in NSDUH, where available. While our results suggest that infoveillance techniques may be useful for tracking emerging drug trends, more research is needed to develop strong methodological approaches to analyzing this data. Issues, such as bias in the data and mismatched units of analysis, need to be addressed. The future of infoveillance research can be further enhanced by a theoretical framework guiding how people talk about, share, and seek out information online.

REFERENCES

Ayers, J. W., Ribisl, K. M., Brownstein, J. S. (2011b).Tracking the rise in popularity of electronic nicotine delivery systems (electronic cigarettes) using search query surveillance. *American Journal and Preventive Medicine, 40*(4),448-453

Ayers, J. W., Ribisl, K., Brownstein, J. S. (2011 a).Using search query surveillance to monitor tax avoidance and smoking cessation following the United States' 2009 "SCHIP" cigarette tax increase. *PLoS One;6*(3), e16777.

Chapman, S., McLeod, K., Wakefield, M., Holding, S. (2005). Impact of news of celebrity illness on breast cancer screening: Kylie Minogue's breast cancer diagnosis. Medical Journal of Autralia, 183(5), 247-250.

Eysenbach, G. (2009). Infodemiology and infoveillance: Framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the Internet. *Journal of Medical Internet Research, 11*(1), e11

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature 457*(7232), 1012-1014.

Goel, S., Hofman, J. M., Lanaie, S., Pennock, D. M., and Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences U S A. 107*(41), 17486–17490

Google Insights for Search (2011a). *Analyzing the data: Is the data normalized?* Available at http://www.google.com/support/insights/bin/topic.py?topic=13975. Accessed on June 1, 2011.

Google Insights for Search (2011b). *Analyzing the data: How is the data scaled?* Available at http://www.google.com/support/insights/bin/answer.py?hl=en&answer=87282. Accessed on June 1, 2011.

Google News (2011). Insight for Search help: Analyzing data. Available at http://www.google.com/support/insights/bin/topic.py?topic=13975. Accessed on June 4, 2011.

Kaiser Family Foundation. (2010). Generation M2: Media in the lives of 8- to 18-year-olds. Available at: http://www.kff.org/entmedia/upload/8010.pdf.

Pew Internet and American Life (2011). Get the latest statistics. Available at http://www.pewinternet.org/Data-Tools/Get-The-Latest-Statistics.aspx. Accessed on June 20, 2011.

Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D. (2008). Using Internet searches for influenza surveillance. *Clinical Infectious Diseases, 47*, 1443–1448.

Rowinski, D. (2011, May 18). *Has Twitter eclipsed 300 million users?* Available at http://www.readwriteweb.com/archives/has_twitter_eclipsed_300_million_users.php. Accessed on June 20, 2011.

Smith, A. (2011, June 1). *Twitter Update 2011*. Pew Internet & American Life Project. http://pewinternet.org/ Reports/2011/Twitter-Update-2011.aspx. Accessed on June 20, 2011.

Substance Abuse and Mental Health Services Administration (SAMHSA). (2010a). Results from the 2009 National Survey on Drug Use and Health: Volume I. Summary of National Findings (Office of Applied Studies, NSDUH Series H-38A, HHS Publication No. SMA 10-4856Findings). Rockville, MD.

Substance Abuse and Mental Health Services Administration. (2010b). Results from the 2009 National Survey on Drug Use and Health: Volume II. Technical Appendices and Selected Prevalence Tables (Office of Applied Studies, NSDUH Series H-38B, HHS Publication No. SMA 10-4856Appendices). Rockville, MD.

Twitter blog. (2011, March 14). # *numbers*. Available at http://blog.twitter.com/2011/03/numbers.html. Accessed on June 20, 2011.

Appendix A. Salvia-Related Headline News and Trends in Salvia Searches, Tweets, and Use 2008 to 2011