Published in final edited form as:
Prev Med. 2014 June ; 63: 112–115. doi:10.1016/j.ypmed.2014.01.024.

Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes

Sean D. Young, PhD, MS[1], **Caitlin Rivers, MS**[2], and **Bryan Lewis, PhD**[2]

[1]Department of Family Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA

[2]Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

Abstract

Objective—Recent availability of "big data" might be used to study whether and how sexual risk behaviors are communicated on real-time social networking sites and how data might inform HIV prevention and detection. This study seeks to establish methods of using real-time social networking data for HIV prevention by assessing 1) whether geolocated conversations about HIV risk behaviors can be extracted from social networking data, 2) the prevalence and content of these conversations, and 3) the feasibility of using HIV risk-related real-time social media conversations as a method to detect HIV outcomes.

Methods—In 2012, tweets (N = 553,186,061) were collected online and filtered to include those with HIV risk-related keywords (e.g., sexual behaviors and drug use). Data were merged with AIDSVU data on HIV cases. Negative binomial regressions assessed the relationship between HIV risk tweeting and prevalence by county, controlling for socioeconomic status measures.

Results—Over 9,800 geolocated tweets were extracted and used to create a map displaying the geographical location of HIV-related tweets. There was a significant positive relationship ($p < .01$) between HIV-related tweets and HIV cases.

Conclusion—Results suggest the feasibility of using social networking data as a method for evaluating and detecting HIV risk behaviors and outcomes.

Keywords

social networking; HIV detection; HIV prevention; big data; digital epidemiology

Address correspondence and reprints to: Sean D. Young, Department of Family Medicine, University of California at Los Angeles, 10880 Wilshire Blvd, Suite 1800, Phone: 1-310-794-8530, Fax: 1-310-794-3580, sdyoung@mednet.ucla.edu.

The authors declare that there are no conflicts of interest.

Introduction

Social networking technologies have recently been used for HIV prevention research (Young, 2012, Gold et al., 2011) as tools for recruitment (Sullivan et al., 2011), interventions (Bull et al., 2012, Young et al., 2013a), and mixed-methods research (Young and Jaganath, 2013). Because people sometimes use these technologies to publicly discuss sexual-related attitudes, desires, and behaviors, researchers may be able to use social networking data to understand and detect real-time individual and regional sexual risk behaviors and social norms (Young and Jordan, 2013). An emerging field, known as digital epidemiology, studies how these "big data" can be used to better understand, detect, and address public health problems (Salathe et al., 2012, Aramaki et al., 2011). However, no known research has been conducted on methods for how or whether these data can be used for HIV prevention or detection, making it important to evaluate the feasibility of this approach. Evaluating methods for how to use social media and "big data" in public health and medicine is an important first step in establishing how these data can be used in prevention, detection, and treatment.

For example, millions of social communications from real-time, geographically-linked, social networking sites, such as Twitter, might be used to make inferences about geographical rates of future or recent past engagement in sexual risk behaviors. Twitter, a large and rapidly growing social networking technology, allows participants to send short, public, real-time "tweet" communications (Smith and Brenner, 2012). Twitter provides public access to these data through an advanced programmatic interface (API) (Twitter, 2013). People who intend to or have just engaged in sexual or drug-related behaviors might tweet to their social networks to inform them of their attitudes and behaviors (Walker, 2013, Young et al., 2013b). Researchers may be able to link these Twitter data to real-time incidence data to better understand and detect public health outbreaks. For instance, influenza researchers have compared flu data with tweets related to influenza symptoms and found tweets have been able to detect influenza outbreaks in regions where the tweets occurred, in advance of traditional surveillance methodologies (Aramaki et al., 2011).

HIV researchers could build on this approach by studying whether engagement in sexual risk behaviors could be inferred from tweet content, for example by filtering for keywords that suggest sexual risk and drug use behaviors (i.e., HIV risk behaviors). Because Twitter provides geographical locations (i.e., geolocated data) for some conversations, HIV risk-related tweets can ultimately be mapped alongside incidence rates to determine whether regional rates of HIV-risk conversations on Twitter could be associated with HIV transmission in those regions. However, these topics have not been studied, making it important to evaluate the feasibility of studying whether and how HIV-risk behaviors are communicated using real-time social media and whether these communications could be linked to allow analysis of data on HIV transmission.

This study is designed to evaluate the feasibility of developing methods of using "big data" to understand whether and how HIV and drug risk behaviors are communicated online in real-time and how these data might be used to inform HIV prevention and detection efforts. Specifically, this study seeks to determine 1) whether geolocated conversations about HIV

risk (sexual and drug use) behaviors can be extracted from realtime social networking data, 2) the prevalence and content of these conversations, and 3) the feasibility of using HIV risk-related real-time social media conversations as a method of remote monitoring and detecting HIV transmission.

Methods

This study received exemption from the Virginia Tech Institutional Review Board. Tweets (N = 553,186,061) were collected from Twitter's free Advanced Programming Interface (API) between May 26, 2012 and Dec 09, 2012. We used Twitter's 'garden hose' method of collecting tweets, which provides a random sample of approximately 1% of all tweets. Tweets collected through the garden hose are available in real time; the data are consistently streamed as the tweets are sent through the service. A variety of metadata are available along with the tweet text including the user's language, number of friends and followers (people who subscribe to the user's communications) and time the tweet was sent. Some users also choose to enable a feature that includes the author's location, in the form of a latitude and longitude, to the tweet. Currently approximately 1% of tweets are geolocated. If users enable geolocated data then this information is also provided through the API.

Data were filtered to include only geolocated tweets originating from the United States, limiting the sample to 2,157,260 tweets. Geolocations in the United States were selected and assigned to the state and county levels as Federal Information Processing Standard (FIPS) codes using Geographic Information Systems (GIS) database operations.

A list of words was compiled that was determined to be associated with sexual risk-related attitudes and behaviors, as well as HIV-related substance use (e.g., stimulants and opiates that have been shown to be associated with HIV (Shoptaw, 2006)). These colloquial words and phrases were coded as being suggestive of sex and substance use behaviors, such as "sex" and "get high." A tweet was classified as a sexual or drug risk-related tweet if it contained one or more risk-related words. Sex and drug risk-related tweets were combined to create an overall category of HIV-related tweets. We created an algorithm that searched the data we collected from Twitter and retrieved tweets with at least one keyword. All words were stemmed and converted to lowercase, and punctuation was removed. Stemming is the removal of suffixes, so that 'waits', 'waited', 'waiting', etc. all become 'wait'. A sample of the filtered tweets was manually checked to ensure they were accurately related to HIV risk behavior. The text of each tweet was processed to maximize sensitivity and specificity of content identification by filtering out tweets that contained co-occurring words that were not associated with HIV risk behaviors (such as removing tweets if "coke" included references to the drink instead of the drug). Based on these results, the list of words in the algorithm was refined to improve the accuracy of the tweets as being related to sexual risk. This process was repeated one time (Figure 1).

No national data were available for use on HIV transmission or incidence. HIV prevalence data were extracted from aidsvu.org, which provides county-level data of HIV/AIDS cases from 2009. The AIDSVU database also includes county data on socio-economic status measures, such as median income, percent living in poverty, percent with a high school

education, and GINI index. The GINI index is a measure of wealth inequality, for which a value of 0 represents complete equality, and a value of 1 represents a circumstance where one person has all of the wealth. A number of states (North Dakota, South Dakota, Vermont, District of Columbia, Hawaii, Alaska, Maryland, and Massachusetts) do not have publicly available HIV data and were therefore excluded from analysis.

Analysis

Counts of HIV-related tweets were tallied from each county and merged with HIV data from aidsvu.org (http://aidsvu.org/about-aidsvu/overview) to create a table with county-level data for analyses. Descriptive statistics for tweet metadata were calculated for sex risk and drug risk-related tweet categories, as well as for the overall demographics of Twitter users sending tweets.

Univariate regressions assessed associations between the proportion of sex, stimulant drug use, and HIV-related (combined sex and drug) tweets and number of HIV cases in that county. The proportion is the count of tweets in that county over the sum of the number of overall tweets. Negative binomial multiple regression assessed the relationship between the proportion of HIV-related tweets from each county and HIV prevalence, percent living in poverty, percent uninsured, percent with a high school education, and the GINI index for each county as covariates. The model includes an offset of the number of people living in that county to adjust for population.

Results

The majority of geolocated tweets, including general as well as HIV-risk related tweets, were sent from California (9.4%), Texas (9.0%), New York (5.7%), and Florida (5.4%). District of Columbia, Delaware, Maryland and Mississippi tweeted the most overall per capita (Table 1).

The algorithm collected 8,538 sexual risk-related tweets and 1,342 stimulant drug use-related tweets, totaling 9,880 HIV-related tweets. District of Columbia, Delaware, Louisiana, and South Carolina sent the largest raw number of HIV risk-related tweets per capita. Utah, North Dakota, and Nevada had the highest per capita rate of HIV-related tweets per overall rates of tweets (see Figure 2).

Results from the univariate analysis showed a significant positive relationship between the proportion of sex risk-related tweets and HIV prevalence at the county level (Coef = 256, $p < .0001$), and the proportion of drug risk-related tweets and HIV prevalence (Coef = 159, $p < .0001$) at the county-level. We found a significant positive relationship between the combined (sex and drug) HIV risk-related category of tweets and county HIV prevalence (Coef = 254, $p < .0001$).

Results from the multivariate regression showed a significant positive relationship between the combined HIV-risk related tweets within a county and county-level HIV prevalence, with percent living in poverty, percent uninsured, percent with a high school education, the GINI index as covariates (see Table 2).

Discussion

This study provides the first set of evidence for how real-time social media data might be used for extracting, detecting, and remote monitoring of health-related attitudes and behaviors. Results suggest the feasibility of using data from real-time social networking technologies to identify HIV risk-related communications, geographically map the location of those conversations, and link them to national HIV outcomes data for additional analyses. Further, tweets that implied HIV-risk behaviors were associated with county-level HIV prevalence, controlling for the overall number of tweets, as well as county-level socio-economic status measures. This study is important because it not only provides support for use of “big data” and information on where and how people are communicating about HIV-risk behaviors online, but because it also provides support for a method of testing whether these data can be used for HIV surveillance. Because of the growing amount of social media data, researchers and public health departments will soon be able to build upon these methods to more accurately monitor and detect health behaviors and disease outbreaks.

Influenza and computer science researchers studying digital epidemiology have already shown that people's tweets can be used to detect actual influenza outbreaks (Aramaki et al., 2011, Culotta, 2010, Chew and Eysenbach, 2010, Lampos et al., 2010). For example, Aramaki et al., extracted tweets and used an approach (similar to the method in this study) to filter tweets that were associated with influenza reports in Japan (Aramaki et al., 2011). When comparing tweets to actual Japanese influenza reports, they found up to a .97 correlation. Influenza like illness data is readily available; case counts are released weekly. Such highly resolved data are not available for HIV, creating a need for more frequent updates in public health and HIV data. The present results suggest that it may be possible to use social networking data to detect real-time HIV transmission if those data are accessible. Results also suggest a potential for public health departments to use social networking data to identify real-time changes in sexual risk behaviors, by location.

This study is limited by a number of factors, most of which relate to lack of access to timely HIV case reports. Instead of assessing the relationship between real-time HIV-related communications and HIV transmission, recent infections, or HIV risk behaviors, data were available only for HIV prevalence from 2009, meaning that instead of detecting HIV outcomes, the resulting relationship between HIV and HIV-related tweets might have resulted from participants living in regions that already had higher HIV prevalence. However, results on the relationship between HIV-related tweets and HIV prevalence were presented more as a demonstration of the method than to demonstrate a predictive causal link between HIV-related tweets and available data on HIV outcomes. Nonetheless, the presence of HIV risk behaviors in areas with high levels of HIV prevalence present opportunities to monitor and possibly interrupt HIV transmission. The present study is a feasibility study primarily designed to test whether real-time social media communications that suggest HIV risk behaviors can be extracted and used as a method for HIV prevention and surveillance. After finding these methods to be feasible, future research can use these methods to explore the relationship between real-time communication and more recent, or even future, HIV transmission cases. Future research can also collect data over a longer time horizon than conducted in this study in order to more effectively evaluate these methods.

Another limitation of this study is that we do not know individual and population-level risk factors associated with the data or have the ability to compare geo-located participant data to non-geolocated data to better understand the participant sample characteristics. It is possible that the Twitter users who provided data for this study were not at high-risk for HIV, and therefore understanding their HIV-risk communications would be less relevant. However, minority populations at high-risk for HIV, such as African Americans and Latinos, are more likely to use social networking technologies than White individuals, making it likely that our sample includes at-risk populations (Smith, 2011). Further, the present analysis controlled for socio-economic status county-level measures such as income and education.

Importantly, HIV-risk communications in each county were associated with higher levels of HIV prevalence. This means that whether or not data are collected from users who are specifically at high-risk, the presence of HIV risk communications on Twitter suggests an association with HIV and provides an important tool for county-level research and interventions. Future research can help to explore how to use geographical data on Twitter for population-focused HIV prevention interventions. Further, a primitive algorithm was used for identifying tweets suggesting HIV risk behaviors. Future research may use more advanced techniques to identify relevant content. This study was conducted to evaluate the feasibility of this novel approach and to initiate methods in this area that can help to create a refined process for identifying HIV risk-related tweets.

Finally, HIV infection is extremely nuanced, affected by biological, medical, and social/behavioral factors. We therefore do not claim that the current form of this analysis can be used to predict HIV behaviors or infection. However, we believe that in the near future, advances in technologies will be available that can be used to more easily and accurately measure and transmit biological data, such as data on infection. We believe that the present methods of using social media data will be able to be enhanced and modeled alongside new and more precise forms of bio-behavioral data to more effectively predict behavior and disease.

Conclusion

Results from this study suggest that it is feasible to use real-time social networking technologies to identify HIV risk-related communications, geographically map the location of those conversations, link them to national HIV outcomes data for additional analyses, and that these data were associated with county-level HIV prevalence. This study was designed to provide a call for future research to understand the potential cost-effectiveness of this approach and to refine methods of using real-time social media data for HIV and public health prevention and detection.

Acknowledgments

We wish to thank the National Institute of Mental Health (NIMH) for funding this work.

References

Aramaki, E.; Maskawa, S.; Morita, M. Proceedings of the Conference on Empirical Methods in Natural Language Processing. Edinburgh, United Kingdom: Association for Computational Linguistics; 2011. Twitter catches the flu: detecting influenza epidemics using Twitter.

Bull SS, Levine DK, Black SR, Schmiege SJ, Santelli J. A Social Media Delivered Sexual Health Intervention: A Cluster Randomized Controlled Trial. American Journal of Preventive Medicine. 2012; 43:467–474. [PubMed: 23079168]

Chew C, Eysenbach G. Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. PLoS ONE. 2010; 5:e14118. [PubMed: 21124761]

Culotta, A. Proceedings of the First Workshop on Social Media Analytics. Washington D.C., District of Columbia: ACM; 2010. Towards detecting influenza epidemics by analyzing Twitter messages.

Gold J, Pedrana AE, Sacks-Davis R, Hellard ME, Chang S, Howard S, Keogh L, Hocking Js, Stoove Ma. A systematic examination of the use of online social networking sites for sexual health promotion. BMC. 2011; 11:583.

Lampos V, De Bie T, Cristianini N, Balcázar J, Bonchi F, Gionis A, Sebag M. Flu Detector - Tracking Epidemics on Twitter. 2010

Machine Learning and Knowledge Discovery in Databases. Springer; Berlin/Heidelberg:

Salathe M, Bengtsson L, Bodnar TJ, Brewer DD, Brownstein JS, Buckee C, Campbell EM, Cattuto C, Khandelwal S, Mabry PL, Vespignani A. Digital Epidemiology. PLoS Comput Biol. 2012; 8:e1002616. [PubMed: 22844241]

Shoptaw S. Methamphetamine use in urban gay and bisexual populations. Topics in HIV Medicine. 2006; 14:84–87. [PubMed: 16835463]

Smith, A. Who's on what: Social media trends among communities of color. In: PROJECT, PIAL. , editor. Race and Ethnicity: Social Networking. California: Immunization Coalition; 2011.

Smith A, Brenner J. Twitter use 2012. Pew Internet and American Life Project. 2012

Sullivan PS, Khosropour Cm, Luisi N, Amsden M, Coggia T, Wingood Gm, Diclemente Rj. Bias in online recruitment and retention of racial and ethnic minority men who have sex with men. J Med Internet Res. 2011; 13:e38. [PubMed: 21571632]

Twitter. Twitter. 2013

Walker P. Police commissioner backs teenage adviser after ‘youthful boasting’ tweets. The Guardian. 2013

Young SD. Recommended Guidelines on Using Social Networking Technologies for HIV Prevention Research. AIDS and Behavior. 2012; 16:1743–1745. [PubMed: 22821067]

Young SD, Cumberland W, Sung-Jae L, Jaganath D, Szekeres G, Coates T. Social networking technologies as emerging tools for HIV prevention: A Randomized Controlled Trial. Annals of Internal Medicine. 2013a; 159:318–324. [PubMed: 24026317]

Young SD, Jaganath D. Online Social Networking for HIV Education and Prevention: A Mixed Methods Analysis. Sexually Transmitted Diseases. 2013; 40:162–7. [PubMed: 23324979]

Young SD, Jordan A. The Influence of Social Networking Photos on Perceived Social Norms and Sexual Health Behaviors. Cyberpsychology, behavior, and social networking. 2013; 16:243–7.

Young SD, Szekeres G, Coates T. The relationship between online social networking and sexual risk behaviors among men who have sex with men (MSM). PLoS ONE. 2013b; 8:e62271. [PubMed: 23658716]

Research Highlights

- We collected 553,186,061 tweets
- We filtered tweets by whether they suggested HIV-related risk behaviors (N = 9,880)
- We presented a visual map of the location of these HIV-related tweets
- We found a significant positive County-level relationship between HIV tweets and HIV prevalence
- We established the feasibility of this method to study HIV-related outcomes

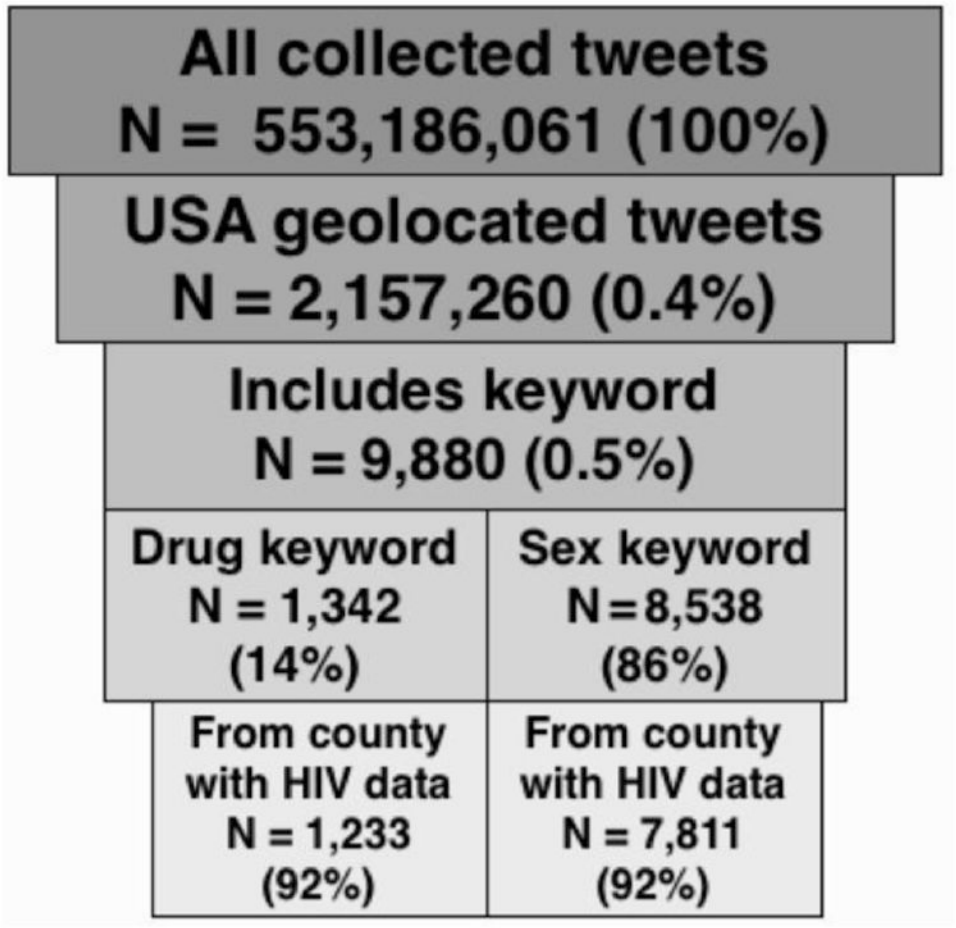

Figure 1. Flowchart of Tweets, USA, 2012

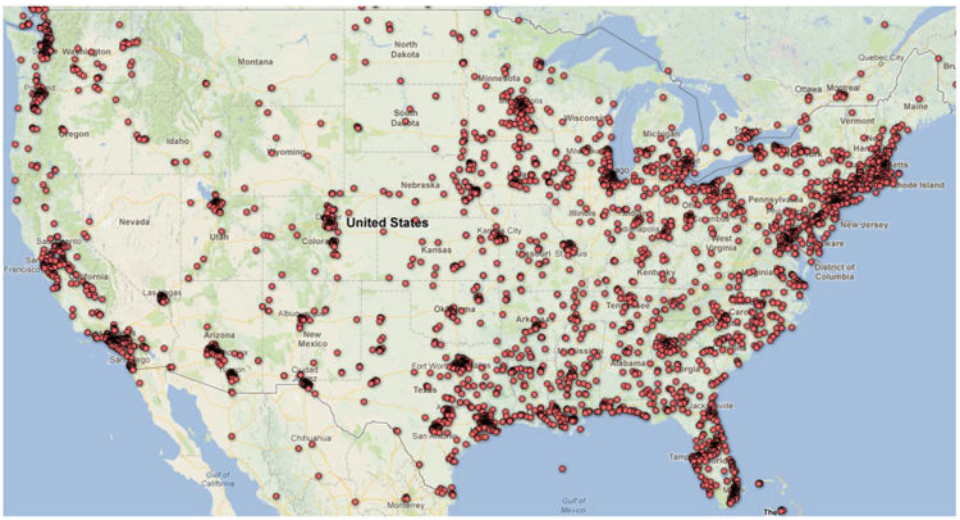

Figure 2.
Map of HIV risk-related geolocated tweets in the United States, 2012.

Table 1

Demographic information about tweets (N = 553,186,061) collected, 2012, United States.

	Sex-related Tweets		Drug-related Tweets		USA Geolocated		All tweets	
	n	**%**	**n**	**%**	**n**	**%**	**N**	**%**
Total	8,538	100.0%	1,342	100.0%	2,157,260	100.0%	553,186,061	100.0
Geo-enabled	8,538	100.0%	1,342	100.0%	2,157,260	100.0%	176,749,975	32.0%
English	8,473	99.2%	1,332	99.3%	2,134,126	98.9%	321,685,106	58.2%
Location listed in tweet	6,577	77.0%	1,049	78.2%	1,626,204	75.4%	385,681,917	69.7%
Geolocation								
California	808	9.5%	182	13.6%	201,887	9.4%	N/A	N/A
Texas	799	9.4%	159	11.8%	193,390	9.0%	N/A	N/A
New York	498	5.8%	111	8.3%	122,730	5.7%	N/A	N/A
Florida	492	5.8%	99	7.4%	117,295	5.4%	N/A	N/A

Prev Med. Author manuscript; available in PMC 2015 June 01.

Table 2

Multivariate analysis of factors associated with county HIV prevalence, United States, 2012

The model includes an offset of the number of people living in that county to adjust for population.

	Coefficient	Standard Error	*p*-value
Proportion of HIV-related tweets	265.0	12.4	<.0001
Percent living in poverty	2.1	0.4	<.0001
GINI index	4.6	0.6	<.0001
Percent without health insurance	1.3	0.4	<.01
Percent with a high school education	-1.1	-3.1	<.01