Preventive Medicine 63 (2014) 112–115

Contents lists available at ScienceDirect

Preventive Medicine

journal homepage: www.elsevier.com/locate/ypmed

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

Sean D. Young [a,*], Caitlin Rivers [b], Bryan Lewis [b]

[a] Department of Family Medicine, David Geffen School of Medicine, University of California, Los Angeles, Los Angeles, CA, USA
[b] Network Dynamics and Simulation Science Laboratory, Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA, USA

ARTICLE INFO

Available online 8 February 2014

Keywords:
Social networking
HIV detection
HIV prevention
Big data
Digital epidemiology

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 9800 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 Human immunodeficiency virus (HIV) risk behaviors and outcomes.

Introduction

Social networking technologies have recently been used for HIV prevention research (Gold et al., 2011; Young, 2012) 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 on 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 Programming 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,

* Corresponding author at: Center for Digital Behavior, Department of Family Medicine, University of California at Los Angeles, 10880 Wilshire Blvd, Suite 1800, USA. Fax: +1 310 794 3580.
E-mail address: sdyoung@mednet.ucla.edu (S.D. Young).

0091-7435/$ – see front matter © 2014 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ypmed.2014.01.024

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 real-time 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 System (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 that 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 (Fig. 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.

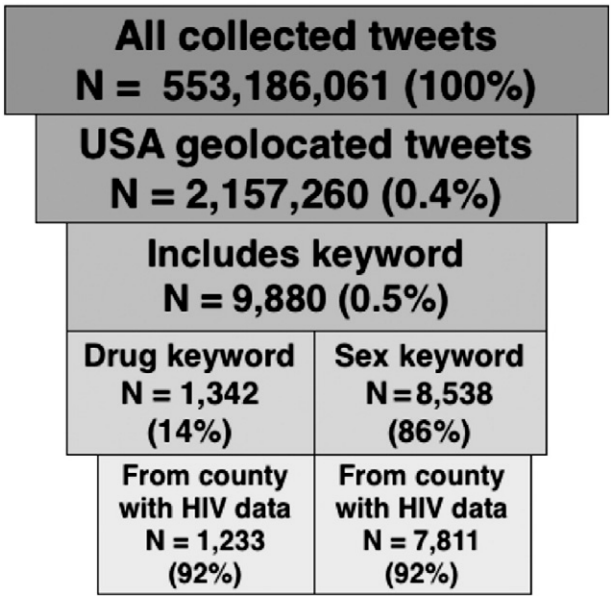

Fig. 1. Flowchart of tweets, USA, 2012.

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 8538 sexual risk-related tweets and 1342 stimulant drug use-related tweets, totaling 9880 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 Fig. 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, $p < .0001$).