Preventive Medicine 109 (2018) 58–61

Contents lists available at ScienceDirect

Preventive Medicine

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

Using social media as a tool to predict syphilis

Sean D. Young[a,*], Neil Mercer[b], Robert E. Weiss[b], Elizabeth A. Torrone[c], Sevgi O. Aral[c]

[a] Department of Family Medicine, University of California, Los Angeles, California, USA
[b] Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles, CA, USA
[c] Division of STD Prevention, Centers for Disease Control and Prevention, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Social media
Twitter
Syphilis

ABSTRACT

Syphilis rates have been rapidly rising in the United States. New technologies, such as social media, might be used to anticipate and prevent the spread of disease. Because social media data collection is easy and inexpensive, integration of social media data into syphilis surveillance may be a cost-effective surveillance strategy, especially in low-resource regions. People are increasingly using social media to discuss health-related issues, such as sexual risk behaviors, allowing social media to be a potential tool for public health and medical research. This study mined Twitter data to assess whether social media could be used to predict syphilis cases in 2013 based on 2012 data. We collected 2012 and 2013 county-level primary and secondary (P&S) and early latent syphilis cases reported to the Center for Disease Control and Prevention, along with > 8500 geolocated tweets in the United States that were filtered to include sexual risk-related keywords, including colloquial terms for intercourse. We assessed the relationship between syphilis-related tweets and actual case reports by county, controlling for socioeconomic indicators and prior year syphilis cases. We found a significant positive relationship between tweets and cases of P&S and early latent syphilis. This study shows that social media may be an additional tool to enhance syphilis prediction and surveillance.

Key points

Question: Can social media be used to predict syphilis?

Findings: We found a significant positive relationship between sexual risk–related tweets and cases of primary and secondary (P&S) and early latent syphilis, independent of the number of syphilis cases in the prior year.

Meaning: Social media is a potential tool to predict syphilis cases; further work is needed to investigate the utility of using social media data to enhance syphilis surveillance. Because social media data collection is easy and inexpensive, integration of social media data into syphilis surveillance may be a cost-effective surveillance strategy, especially in low-resource regions.

1. Introduction

Syphilis, a sexually transmitted disease caused by the bacterium *Treponema pallidum,* is a growing public health concern. Untreated or inadequately treated infection can lead to long-term sequelae, including ocular complications and damage to the internal organs, and can facilitate HIV transmission (Sparling, n.d.). Although rates of syphilis hit historic lows in 2001, rates have steadily increased since then, with a 19% increase occurring from 2014 to 2015 (STDs on the Rise Press Release, 2015). Traditionally, researchers and public health professionals have relied on surveillance strategies such as case reporting and patient interviews to monitor trends in reports of syphilis (Centers for Disease Control and Prevention, 2003). However, a major limitation of these strategies is that they require extensive time and resources to gather and analyze data, and may suffer from self-report biases. To address the growing syphilis epidemic, additional tools are needed.

A large and increasing number of people have been using social media sites over the past decade (*Street 1615 L., NW, Washington S 800, Inquiries D 20036 202 419 4300 | M 202 419 4349 | F 202 419 4372 | M. Social Networking Fact Sheet [Internet]*, 2013). Some social media users publicly discuss sexual risk–related attitudes, desires, and behaviors, which provides an opportunity for researchers to collect large amounts of data on sexual risk behaviors without the costs associated with traditional study techniques. Social media has been shown to be a feasible tool for detecting HIV case diagnoses and stress levels among students (Liu et al., 2017). For example, in 2014, Young and colleagues found a strong, positive correlation between county-level HIV incidence rates and HIV risk related tweets, suggesting the feasibility of using social

* Corresponding author at: University of California, Los Angeles, University of California Institute for Prediction Technology, Department of Family Medicine, 10880 Wilshire Blvd., Ste. 1800, Los Angeles, CA 90024, USA.
E-mail addresses: Sdyoung@mednet.ucla.edu (S.D. Young), Robweiss@ucla.edu (R.E. Weiss), Etorrone@cdc.gov (E.A. Torrone), Saral@cdc.gov (S.O. Aral).

https://doi.org/10.1016/j.ypmed.2017.12.016
Received 1 August 2017; Received in revised form 15 December 2017; Accepted 18 December 2017
Available online 24 December 2017
0091-7435/

media in detecting HIV (Young et al., 2014). Because people are using these sites to share health-related information, researchers have suggested incorporating social media data to monitor public health events, including influenza rates (Aramaki et al., 2011; Chew and Eysenbach, 2010) and cardiovascular disease (Eichstaedt, 2016).

Public health organizations, including the Centers for Disease Control and Prevention (CDC), have expressed interest in evaluating whether social media data might be used to address sexually transmitted diseases (STDs), such as syphilis (http://www.cdc.gov/socialmedia/tools/guidelines/index.html, n.d.). If it were possible to predict STD trends, organizations could develop targeted intervention programs and more efficiently allocate resources toward high-risk geographic areas.

This study was designed to explore whether social media data might be used to predict trends in primary and secondary syphilis and early latent syphilis. More specifically, we sought to identify associations between sexual risk-related social media (i.e., Twitter) data and county-level syphilis cases that would be reported the following year.

2. Methods

Syphilis is a nationally notifiable condition in the United States (Centers for Disease Control and Prevention, n.d.). We reviewed weekly county-level syphilis case data reported to the CDC by all 50 states and Washington D.C. from 2012 to 2013. Weekly data were combined to give annual number of cases. Case reports were restricted to primary and secondary (P&S) syphilis cases (the earliest and most transmissible stages of syphilis) and early latent syphilis cases (the stage capturing non-P&S syphilis cases that likely occurred in the past 12 months); these stages of syphilis most closely represent incident infections. We included a number of covariates, including socioeconomic status, GINI index, previous year syphilis cases, population size, and education levels by geography. These covariates were added based on either previous research on typical associations with sexually transmitted and/or other related diseases as well as based on our own intuitions of possible confounding variables. Socioeconomic variables for all counties were obtained from the U.S. Census Bureau for 2012 to 2014. These covariates included county-level data on the percentage of residents living in poverty, the percentage of residents without health insurance, the percentage of residents with a high school education, and the GINI index.

The GINI index, a measure of wealth inequality ranging from zero (equal distribution of wealth) to one (one person has all of the wealth), was used to estimate the amount of wealth disparity in the United States. The GINI index is available for all 50 states, but county-level data were available only for 814 counties. The GINI index was unavailable for rural counties so the state-level GINI index was substituted in its place. As a county's prior syphilis rate likely predicts future syphilis rates, we also included 2012 syphilis data as a covariate to determine whether prior year twitter data would continue to predict future cases independent of prior syphilis burden.

Twitter data were taken from a previous study that identified tweets associated with sexual risk behaviors, drug use, and HIV (Young et al., 2014). In that study, tweets were collected using Twitter's free advanced programming interface between May 26, 2012, and December 9, 2012. Metadata collected along with the tweet text included the user's primary language, number of followers, and time the tweet was sent. Some users also enabled a geolocation feature that disclosed the author's location in the form of latitude and longitude. Currently, approximately 2% of all users provide access to their geocoded information. We filtered the data to include only sex-related tweets and geolocated tweets originating from the United States, which resulted in a sample of 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 (Fig. 1, Table 1).

Because syphilis is largely dependent on sexual behaviors, a list of words (for example, sex, fuck, dick, cock, suck) was compiled that was determined to be associated with sexual risk-related attitudes and behaviors. A tweet was classified as syphilis risk-related if it contained one or more words or colloquial terms related to sexual intercourse.

We created an algorithm that searched the collected data to identify tweets with at least one keyword. All words were stemmed (i.e., suffixes were eliminated), converted to lowercase text, and punctuation was removed. Then, a sample of the filtered tweets was manually checked to ensure they were accurately related to risk-related 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 risk behaviors. Based on the results, the list of words in the algorithm was refined to improve the accuracy of the tweets as being risk-related.

The time and location of each risk-related tweet were noted and the total number of risk-related tweets from each county was calculated. The number of tweets per county ranged from 0 to 300 with a mean value of 2.38; three transformations were used to adjust for right skewness in the number of tweets. The original data analysis considered the proportion of tweets that were risk-related, but access to the total number of tweets per county was unavailable for the syphilis analysis. Risk-related tweets per 100,000 people was substituted for proportion of tweets, and is referred to as *per-capita tweets*. In addition to the per-capita transformation, a log transformation and square-root transformation—traditional transformations used for right-skewed data—were also implemented to evaluate the tweets.

To investigate whether Twitter data could be used to predict observed syphilis case report data, a series of negative binomial regressions were run to analyze the relationship between the number of risk-related tweets per 100,000 people in each county, and reported cases of syphilis in 2013, controlling for possible confounders that were added as covariates.

3. Results

The dataset included 8538 geolocated tweets from the United States that we identified as being related to sexual risk behaviors in 2012. In 2013, there was at least one case of P&S syphilis reported in 1670 counties and at least one case of early latent syphilis reported in 1594 of the 3142 county and county-equivalents (e.g., independent cities that do not belong to a county). As expected, we found that counties with a higher number of syphilis cases in the prior year, 2012, were associated with a 0.6% increase in the 2013 P&S syphilis incidence and a 0.4% increase in early latent Syphilis ($p < 0.0001$).

As hypothesized, counties with a higher number of risk-related tweets from 2012 were significantly associated with a 2.7% increase in P&S syphilis and a 3.6% increase in early latent syphilis cases in 2013 ($p < 0.0001$). These results were constant across all models run in our analyses, suggesting a relationship between syphilis and risk-related tweets. Risk-related tweets from the previous year predicted syphilis cases, independent of the previous year's syphilis case data and county-level socioeconomic factors.

In the P&S syphilis analysis, a one unit increase of P&S cases from the previous year above the mean led to a 0.6% increase in P&S syphilis cases in the following year, a one unit increase in risk-related tweets (i.e., an increase of one tweet) above the mean led to a 3% increase in P&S syphilis cases, a 1% increase in poverty above the mean led to a 4% increase in P&S syphilis cases, and a 0.10 point increase in GINI led to a 30% increase in P&S cases. In the early latent syphilis analysis, a one unit increase of the early latent syphilis cases from the previous year above the mean led to a 0.3% increase in early latent cases, a one unit increase in risk-related tweets above the mean led to a 4% increase in early latent cases, a 1% increase in poverty above the mean led to a 4% increase in early latent cases, and a 1% increase in the people without health insurance above the mean led to a 4% increase in early latent cases (Table 2). Table 2 displays two separate multivariate regression

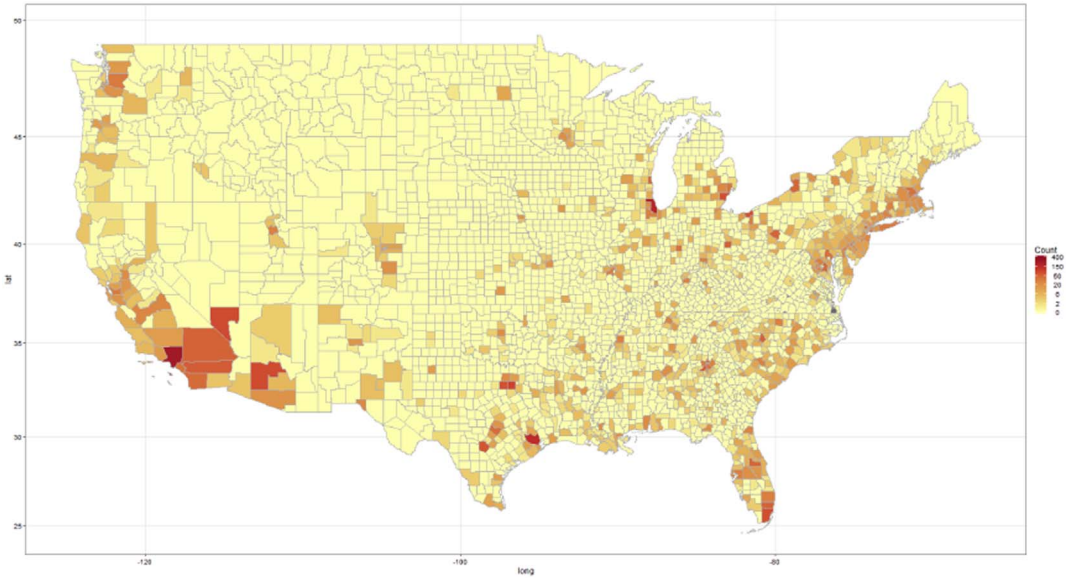

Fig. 1. Risk-related tweets from May 2012 to December 2012.

Table 1
Descriptive statistics.

	Median	Mean	SD	Minimum	Maximum
P&S syphilis cases 2012	0.00	4.87	32.76	0	943
P&S syphilis cases 2013	0.00	5.42	35.42	0	1095
Early latent syphilis cases 2012	0.00	4.52	35.52	0	1345
Early latent syphilis cases 2013	0.00	5.28	39.84	0	1401
Risk-related tweets	0.00	2.38	10.16	0	292
Population 2013	26,084	102,167	326,147	103	10,045,180
GINI 2013	0.46	0.46	0.02	0	1
Poverty percent 2013	16.30	17.23	6.58	4	55
Percent uninsured 2013	17.20	17.43	5.40	3	40
Percent high school 2013	81.40	80.37	8.74	25	100

Table 2
Negative binomial regression results for 2013.

	Coefficient	Standard error	*p*
Primary and secondary syphilis			
P&S syphilis cases in 2012	0.006	0.001	< 0.0001
Risk-related tweets	0.027	0.004	< 0.0001
Percent living in poverty	0.038	0.006	< 0.0001
GINI index	2.611	1.153	0.02
Percent without health insurance	0.002	0.007	0.83
Percent with a high school education	0.004	0.004	0.36
Early latent syphilis			
Early latent syphilis cases in 2012	0.004	0.001	< 0.0001
Risk-related tweets	0.035	0.004	< 0.0001
Percent living in poverty	0.041	0.006	< 0.0001
GINI index	1.509	1.210	0.21
Percent without health insurance	0.043	0.007	< 0.0001
Percent with a high school education	0.009	0.004	0.05

models: one for P&S Syphilis and the other for early latent syphilis. The table includes all predictors, including ones that were not statistically significant.

4. Discussion

Researchers have suggested that social media may be a new tool that could be used to predict public health outcomes, such as syphilis cases, and supplement existing surveillance tools (Young et al., 2014). Previous studies have used similar methods to find associations between risk behaviors reported on social media and HIV prevalence (Young et al., 2014). To our knowledge, this is the first study to explore the association between risk behaviors and syphilis. We found that reported syphilis cases were predicted by the number of risk-related tweets per county in the previous year, suggesting that further research be conducted to explore whether and how social media might be used to forecast trends in syphilis, including geographical and temporal variations. Theoretically, researchers could apply this methodology of mining social media for risk behaviors to numerous diseases and health events.

Epidemiologists and health departments have already begun to use social data to monitor public health outbreaks. For example, Aramaki and colleagues extracted and filtered tweets that were associated with influenza reports in Japan (Aramaki et al., 2011). When the tweets were compared to actual Japanese influenza reports, up to a 0.97 correlation was found. Elsewhere, in a study of cardiac incidents, (Eichstaedt, 2016) researchers evaluated 100 million tweets from 1300 U.S. counties. The language in the tweets was parsed into word clouds that reflected "risky" or "protective" language. Similar to the methods used in this study, algorithms were created that compared the sentiment of the tweets from each county with CDC data on causes of death. Counties whose tweets expressed more negative emotion had more heart disease–related deaths compared with counties that featured tweets with more protective language. The present study provides more support for the claim that social media data might be useful as a tool for forecasting regional disease trends and warrants additional research on this topic.

In this era of limited resources, it is important to explore all possible methods that might be used to improve public health. Given that initial findings from this study and others support that social media data might be used as a public health monitoring and/or prediction tool, additional research is warranted on this topic. If social media monitoring models were found to be accurate and effective, then public health organizations could be able to use results from these models to intervene using existing syphilis intervention methods. For example, public health organizations might use geo-targeted social media data to help facilitate increased syphilis partner notification, screening in jails and correctional facilities, and to encourage providers to screen according to recommendations and guidelines.

This study was limited by a number of factors. First, although a large and increasing number of people use social media, including

Twitter, these data are a biased representation of the general population. Mitigating this issue somewhat is the fact that young people are more likely to use Twitter, and are more likely to be diagnosed with syphilis (STDs on the Rise Press Release, 2015; Greenwood et al., 2016). Second, this analysis only reviewed data from 2012 and 2013: a longitudinal analysis would be needed before public health organizations invested resources to develop formal mechanisms based on this analysis. In addition, the majority of tweets could not be geolocated, and areas with a high number of syphilis cases may already have public health messaging programs that use social media. For example, although we checked the content of tweets to verify that the majority of tweets discussed content related to sexual risk, many were also related to prevention of STDs that may have been created by public health organizations. However, we still believe the association identified in this analysis is important and warrants additional exploration as it occurs regardless of the limitations of social media content and location of tweets. Additional limitations were that syphilis was defined using reported cases, which likely underestimated incidence; (Satterwhite et al., 2013) and, as in all non-randomized controlled studies, there may be additional confounders that we did not include.

Finally, it is important to address the ethical implications of this study. While the tweets we analyzed did contain personal information (i.e., location and username) and admissions of risky behaviors, it is important to recognize that the Twitter profiles were all public. In the future, if researchers wish to use data from private social media profiles as well as public, social media platform hosts will have to consider the privacy of their users when granting permissions to researchers.

5. Conclusion

Surveillance for syphilis, including routine case reporting, is the backbone of prevention and control activities. However, new tools are needed to address the limitations of existing methods such as lag time in reporting. Social media data are captured in a natural setting and publicly and freely available in real-time, helping to address some of the limitations of existing tools. This exploratory study suggests that public health organizations may be able to incorporate social media data along with existing methods to more effectively intervene and prevent the spread of syphilis. In areas with robust case reporting, further investigation into associations between social media and real time observed case reports will help to refine prediction models and investigate the utility of social media data. In the absence of syphilis monitoring methods, such as in certain global low-income regions, this study suggests social media data is an inexpensive tool that might be used to estimate case reporting.

Conflict of interest

None to declare.

Funding

This work was supported by support from the National Institute of Mental Health (NIMH) grant 5R01MH106415 (Young); the National Institute of Allergy and Infectious Diseases grant R56 (Young); the Center for HIV Identification, Prevention, and Treatment NIMH grant P30MH058107 (Weiss and Young); and the UCLA Center for AIDS Research grant 5P30AI028697, Core H (Weiss and Young).

References

CDC Social Media Tools, Guidelines & Best Practices | Social Media | CDC [Internet]. [cited 2016 Oct 31]. Available from: http://www.cdc.gov/socialmedia/tools/guidelines/index.html.

Aramaki, E., Maskawa, S., Morita, M., 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing [Internet]. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1568–1576. [cited 2016 Oct 31]. (EMNLP '11). Available from: http://dl.acm.org/citation.cfm?id=2145432.2145600.

Centers for Disease Control and Prevention, 2003. Recommendations for Public Health Surveillance of Syphilis in the United States [Internet]. Atlanta, GA. Available from: https://www.cdc.gov/std/syphsurvreco.pdf.

Centers for Disease Control and Prevention 2016 Nationally Notifiable Conditions [Internet]. [cited 2016 Nov 22]. Available from: https://wwwn.cdc.gov/nndss/conditions/notifiable/2016/.

Chew, C., Eysenbach, G., 2010 Nov 29. Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PLoS One 5 (11), e14118. http://dx.doi.org/10.1371/journal.pone.0014118.

Eichstaedt, J.C., 2016 Mar 1. Status update: stressed, angry, at risk? Sci. Am. Mind. 27 (2), 62–67.

Greenwood, S., Andrew, Perrin, Duggan, M., 2016. Social Media Update 2016 [Internet]. Pew Research Center: Internet, Science & Tech. [cited 2016 Nov 21]. Available from: http://www.pewinternet.org/2016/11/11/social-media-update-2016/.

Liu, S., Zhu, M., Yu, D.J., Rasin, A., Young, S.D., 2017. Using real-time social media technologies to monitor levels of perceived stress and emotional state in college students: a Web-based questionnaire study. JMIR Ment. Health. 4. http://dx.doi.org/10.2196/mental.5626.

Satterwhite, C.L., Torrone, E., Meites, E., Dunne, E.F., Mahajan, R., Ocfemia, M.C.B., et al., 2013 Mar. Sexually transmitted infections among US women and men: prevalence and incidence estimates, 2008. Sex. Transm. Dis. 40 (3), 187–193.

Sparling P. Sexually transmitted diseases, Fourth Edition: King Holmes, P. Sparling, Walter Stamm, Peter Piot, Judith Wasserheit, Lawrence Corey, Myron Cohen. In: Sexually Transmitted Diseases [Internet]. Clinical Manifestations of Syphilis. New York: McGraw-Hill; [cited 2016 Nov 22]. p. 661–84. Available from: https://www.amazon.com/Sexually-Transmitted-Diseases-Fourth-Holmes/dp/0071417486

STDs on the Rise Press Release | 2015 | Newsroom | NCHHSTP | CDC [Internet]. [cited 2016 Oct 31]. Available from: http://www.cdc.gov/nchhstp/newsroom/2015/std-surveillance-report-press-release.html

Street 1615 L., NW, Washington S 800, Inquiries D 20036 202 419 4300 | M 202 419 4349 | F 202 419 4372 | M. Social Networking Fact Sheet [Internet]. Pew Research Center: Internet, Science & Tech. 2013 [cited 2016 Oct 31]. Available from: http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/

Young, S.D., Rivers, C., Lewis, B., 2014 Jun. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev. Med. 63, 112–115. http://dx.doi.org/10.1016/j.ypmed.2014.01.024.