AIDS and Behavior
https://doi.org/10.1007/s10461-018-2046-0

ORIGINAL PAPER

An Online Risk Index for the Cross-Sectional Prediction of New HIV Chlamydia, and Gonorrhea Diagnoses Across U.S. Counties and Across Years

Man-pui Sally Chan[1] · Sophie Lohmann[1] · Alex Morales[2] · Chengxiang Zhai[2] · Lyle Ungar[3] · David R. Holtgrave[4] · Dolores Albarracín[1]

Abstract

The present study evaluated the potential use of Twitter data for providing risk indices of STIs. We developed online risk indices (ORIs) based on tweets to predict new HIV, gonorrhea, and chlamydia diagnoses, across U.S. counties and across 5 years. We analyzed over one hundred million tweets from 2009 to 2013 using open-vocabulary techniques and estimated the ORIs for a particular year by entering tweets from the same year into multiple semantic models (one for each year). The ORIs were moderately to strongly associated with the actual rates (.35 < *r*s < .68 for 93% of models), both nationwide and when applied to single states (California, Florida, and New York). Later models were slightly better than older ones at predicting gonorrhea and chlamydia, but not at predicting HIV. The proposed technique using free social media data provides signals of community health at a high temporal and spatial resolution.

Keywords HIV · Chlamydia · Gonorrhea · Social media · Big data

Introduction

Big data methods to analyze social media are opening the door to inexpensive ways of data collection: Social representations and behavioral patterns that used to be accessible only through survey data or direct observation can potentially be inferred by using social media to gauge the temperature of a geographic or online community. These methods may in the future allow us to study social media to understand geographic variability and assess public health needs in different regions. In the meantime, this potential may be assessed by studying the power of social media data to predict disease patterns. For example, mining social media data, localizing those data, and then using them to predict HIV and sexually transmitted infections (STIs) such as gonorrhea or chlamydia is critical to determine whether these data can be used to understand important social and disease patterns.

Since 2014, the Centers for Disease Control and Prevention have reported a steady increase in cases of STIs in the United States, including an 18.5% rate increase in gonorrhea from 2015 to 2016 alone [1]. There is no report of such an increase in HIV, however, HIV remains a significant burden to public health in the U.S. [2]. Despite long-standing efforts to reduce transmission of STIs, an estimated 20 million STIs are diagnosed each year, creating approximately $517 million, $162 million, and $12.6 billion in healthcare costs for chlamydia, gonorrhea, and HIV, respectively [3]. As for any infectious diseases, understanding the local geographic patterns of STIs is essential to predict in which areas people

Man-pui Sally Chan and Sophie Lohmann have contributed equally to this work and are listed in alphabetical order.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10461-018-2046-0) contains supplementary material, which is available to authorized users.

✉ Man-pui Sally Chan
sallycmp@illinois.edu

[1] Department of Psychology, University of Illinois at Urbana-Champaign, Champaign, IL 61820, USA

[2] Department of Computer Science, University of Illinois at Urbana-Champaign, Champaign, USA

[3] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, USA

[4] School of Public Health, Johns Hopkins University, Baltimore, USA

Published online: 09 February 2018

are most at risk of infection through another person. This geographic information can facilitate both treatment and prevention efforts in the areas most affected by the disease. Therefore, a big data approach that taps into social media to rapidly predict in which areas many people are infected may provide important information.

The social media data could correlate with HIV and STI prevalence for two reasons. First, specific words like "HIV" may be associated with prevalence if people pose questions or reveal their status in the public social media. More likely, social media correlates of prevalence will include words that are not about HIV and STIs but that act as indirect indicators of social trends (entertainment, dating, financial struggles, etc.) discussed in social media, which may be related to health in a community. Therefore, social media postings about sports, daily life, nightlife, and a myriad of other issues may indicate epidemiological differences across regions.

Developing a social media index that gauges HIV/STI prevalence in a region is important, and yet in some contexts, no data of broad and timely geographic coverage are available at all. A complete absence of data is not common in the United States, but limited data availability could occur if shifting government priorities reduce investments in science, public health, or marginalized populations [4, 5]. In addition, countries in which surveillance efforts are weak [6], or in which administration suppresses these data or reports unreliable data for marginalized key populations [7], may benefit from having proxy measures of prevalence available for use in the community. Furthermore, promising results in this area will open up a range of possibilities such as developing social media measures for factors that are not commonly surveyed by government agencies on a national level, such as attitudes, homophobia, or intentions to use PrEP in particular regions. Therefore, testing social media indices of risk is an essential first step in advancing a promising measurement area.

Are there population considerations that would support the use of social media to predict patterns of HIV/STI prevalence? Yes–to begin, HIV and STIs are more common in younger populations, which are also heavy users of social media [8–15]. A recent Pew Internet study showed that 86% of 18–29 year olds use social media, as compared to 80% of 30–49 year olds and 64% of 50–64 year olds [16, 17]. Social media users are also diverse in ethnicity. In the U.S., 69% of White, 63% of Black, and 74% of Hispanic people use at least one social media site. Priority populations in the area of HIV/STI also use social media. For example, men who have sex with men continue to bear the burden of HIV infection and are often heavy Internet and social media users. Studies reveal that about one-third of attendants of a gay pride festival had met a sexual partner on the Internet [18]. Opioid-using populations are now receiving attention and populate social media as well. A recent article in a popular media outlet has presented an investigative report on the use of social media in opioid-using populations [19]. Even though social media messages are likely to reflect the social temperature of an area irrespective of who posts these messages, the overrepresentation of young, at-risk populations online makes finding social media proxies for HIV and STIs highly promising.

In addition, social media can provide fast predictions. Social media postings such as text messages can be tracked in real time, which contrasts with the delays in public reporting of STI data in many U.S. counties. For instance, as of February 2018, the most recent US-wide HIV data available on the CDC websites and at AIDSVu.org are new diagnosis rates from 2016, with chlamydia and gonorrhea rates showing similar reporting delays [2]. Such a delay limits community-initiated education programs and prevention campaigns and increases the difficulty of addressing and anticipating health risks as they emerge. Social media data can provide faster insights, potentially contributing to prevention goals. If epidemic patterns change such that new outbreaks cannot be anticipated based on prior rates or traditional, slow-to-change demographic factors, social media data may provide more current insights than the time-lagged surveillance data. In this way, the results of social-media-based analyses can provide information for preventive and control actions.

In this paper, our objective was to create an *Online Risk Index* (ORI), to be used as a signal or proxy of county-level HIV, gonorrhea, and chlamydia rates. Using county as the unit of analysis, we first developed ORIs by linking social media data (i.e., text messages), mined with computerized Natural Language Processing methods, to CDC-provided measures of new HIV/STIs diagnoses data and then evaluated their performance by correlating the ORIs with the actual disease rates. In the present work, these *signals* were obtained using cross-sectional models in which predicted rates of HIV, chlamydia, and gonorrhea for a particular year were obtained based on language-based predictors from the same year.

Social Media in Predictions of Disease and Social Patterns

Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information [20–24]. An emerging field is adopting social media data to understand public health problems such as influenza [25–27], HIV [22, 28–30], and heart disease mortality [31]. Specifically, language features on social media may be used as a signal but are unlikely to be explicit mentions of disease risk in a geographic region, such as a county.

Social media sites with publicly available text data, geographic location information, and sufficient popularity are optimal for the proposed language-based analyses. Twitter, one of the most popular free micro-blogging services, is ideal because of its public nature, precise geolocation information, and high volume. One in five Americans uses Twitter and users mainly write text messages to communicate their thoughts, attitudes, and behaviors in textual form by posting a 140-character message termed *tweet* (and the character limit has been recently relaxed to 280 since mid-2017) [32–36]. Further advantages of using Twitter as a research tool include that the public data are available via authorized application program interfaces (APIs), that part of the data can be mapped to geographic locations [37], and that it is possible to obtain the data in real-time.

Despite the popularity of Twitter, only a few studies have started to use Twitter data to explore the relationship with HIV rates, and most of them have focused on testing whether or not there is an association [22, 29]. For example, Young et al. [22] found positive associations between tweets with sex- or drug-related words and HIV prevalence data. Likewise, Ireland et al. [29] examined the frequencies of action-oriented words (e.g., work and plan) in associating with county-level HIV prevalence rates. In a separate work attempting to predict HIV outbreak, Ireland et al. [30] used a future-oriented dictionary to analyze what kind of topic might predict HIV prevalence rates in U.S. counties. They used tweets from June 2009 to March 2010 [38] and identified two topics that were negatively and positively associated with HIV prevalence rates in counties. However, the overall performance of such a topic model in predicting HIV/STI rates of all U.S. counties was largely unclear. Taken together, previous work has analyzed data from a short time, rather than multiple years as proposed here. No study, to our knowledge, has tackled the prediction analyses by including other STI rates in addition to HIV prevalence rates.

Machine learning methods, such as *topic modeling* and *k-fold cross-validation* techniques, allow the analysis of large and complex datasets while avoiding overfitting. For language data, *topic modeling* can find words that often occur together and group them into semantically related topics. Then, we can find out which of these topics are most predictive of, for example, HIV diagnoses in counties using non-parametric methods. These topic modeling methods are based on probability distributions, allowing us to include tens of thousands of words in the analysis without needing to conduct tens of thousands of significance tests. Because the topics are the final predictors of HIV rates, each word is relevant only insofar as it contributes to a semantic topic of multiple words. Importantly, we follow a standard procedure—*k*-fold cross-validation, which means using out-of-sample prediction. Specifically, we train a topic model based on a random subsample of the data. Due to the size of our dataset, the training model will fit the data extremely well, and thus a more important question is how well the model predicts the remaining *out-of-sample data*. To answer this question, we test the model's performance on the out-of-sample data, and the correlation coefficients between predicted and actual values are then unlikely to be the result of overfitting a model. These techniques focused on the natural language processing of text data, which are nonspecific to Twitter postings and applicable to other sources of text data, such as Facebook data.

Closed Versus Open-Vocabulary Methods and Precedents of Our Approach

Most previous health studies adopted a closed-vocabulary approach using pre-identified lists of words [22, 29, 30, 39]. Such an approach requires prior assumptions about which categories might predict the outcome (e.g., sexual behavior) and which words the community uses to talk about these categories on social media (e.g., slang terms for sexual behavior). Risky behaviors such as condomless sex, however, may not be discussed online as frequently as, for instance, having the flu. Furthermore, people are likely to communicate on social media using informal, constantly changing writing conventions that suit their needs, culture, and idiosyncrasy [40, 41] and that can emerge in ways that researchers and practitioners cannot foresee. These challenges place limits on the use of closed-vocabulary approaches because pre-identified words are unlikely to capture all relevant semantic features that are associated with health risk.

An open-vocabulary approach [42] may identify signals but not offer any explanation of diseases because it allows for flexible and dynamic linguistic patterns to emerge. In this approach, frequent words and clusters of words are connected to the outcome variable in an exploratory, bottom-up fashion, which can be seen as a hallmark of big data methods [43]. These methods are comparable to the study of environmental traces on human behavior as a way of gauging personality. For example, Park et al. [44] used Facebook Wall messages to assess individuals' personality profiles, and Kosinski et al. [45] collected Facebook Likes data to predict dispositional characteristics. The approach is also analogous to psychological testing methods in which the items on a test do not bear direct resemblance with the traits being measured. The Minnesota Multiphasic Personality Inventory (MMPI) is an excellent illustration of this approach. The MMPI includes over one thousand items, and analytical models have been developed to identify patterns of responses across different subscales for diagnosing potential psychological issues [46]. In the case of analyzing Twitter messages, the open-vocabulary approach identifies topics in a language sample through machine learning methods.

In the area of psychology, open-vocabulary analyses of Facebook Wall data (i.e., short and simple text messages) of sixty-six thousand Facebook users have been used to identify predictive models of personality traits based on language use [44]. Using extraversion as an example, the most predictive topic word cloud includes semantically related words, e.g., hanging, love, ready, providing signals to predict but not explain extraversion. Open-vocabulary analyses outperformed the closed-vocabulary models, as reported in other studies [42, 47]. The predicted personality scores of 4824 Facebook users and their self-reported personality scores correlated *r*s. = .35–.47, indicating that the language-based model using the open-vocabulary approach can provide valid personality information.

The Present Research

We used an open-vocabulary analytical framework to build ORIs and assessed their predictive power using out-of-sample validation methods as safeguards against spurious correlations and overfitting. We obtained ORIs from five waves of de-identified Twitter data and provided a test of whether the prediction value of ORIs differs across years. In other words, is it possible to fit a model in 1 year and continue to use it for prediction across subsequent years? If so, this would represent a significant reduction of effort in the application of big data techniques to public health surveillance.

Methods

Overview

This project included Twitter text messages and HIV/STI rates. We collected over 3 billion tweets, identified their U.S. counties of origin, and derived word matrices for each county. Using county as the unit of analysis, we regressed the language features on the HIV/STI rates to train models and used these models to calculate the ORIs of HIV/STIs, which we correlated with the actual new HIV/STI diagnoses from 2009 to 2013. Because we are predicting community health, not individual health, the data do not include any inferences about particular users, and the present study was considered non-human-subject research by the Institutional Review Board of our university. We combined all tweets for each county and processed them in an aggregated way without taking user demographics into account. All results are presented in a de-identified fashion.

Table 1 Number of geo-mapped tweets in 2009–2013

Year	Coordinates	Profile location	Total
2009	6,215,264	41,479,196	47,694,460
2010	2,758,129	25,847,235	28,605,364
2011	246,561	3,966,570	4,213,131
2012	1,991,208	33,355,886	35,347,094
2013	3,011,939	35,782,401	38,794,340
Total	17,850,568	164,284,849	182,135,417

Data Sources

Twitter Data

Twitter provides free access to tweets via application program interfaces (APIs; https://dev.twitter.com/overview/api). We used the *Gardenhose* API to obtain a 10% random sample of all tweets in 2009–2010, and, after this formerly free service transitioned to commercial agreements, we used the *Streaming* API to collect a 1% random sample in 2011–2013. We obtained over 3 billion tweets and retweets (i.e., reposts of other users' messages) posted between June 2009 and December 2013. We included retweets because the content that individuals decide to retweet is also informative about popular topics and conversations within a community. Given county as the unit of analysis, we used the timestamp to exclude tweets not originating from U.S. time zones and combined users' profile location (if available) with each tweet's precise latitude and longitude coordinates (if available) to map tweets to U.S. counties using Geographic Information Systems (GIS) database operations. We discarded other user information in the geo-mapping process, and it located 155 million tweets to U.S. counties (see Table 1). Tweets from the same county were combined into a single file, which was later used as a single unit in the analyses. Thus, the data could not be used to re-identify individual users and no demographics of sources of the tweets are available.

STI Data

County-level new diagnosis rates per 100,000 of HIV, chlamydia, and gonorrhea were obtained from the CDC [2, 48] and AIDSVu (http://aidsvu.org/) [49]. HIV data included only people aged 13 and older. States for which HIV data are not released at the county-level (i.e., Alaska, District of Columbia, South Dakota) were excluded. Data from counties with less than five new HIV diagnoses per year or less than 100 inhabitants are routinely suppressed by the CDC, and were thus not available for analysis.

Analytic Procedures

Mapping Tweets to Counties

We used the location and coordinate information available in the Twitter metadata to map each tweet to a county [37]. This method relies on either the coordinate information attached to a tweet (latitude, longitude; available only in 8% of all tweets) or the free-response location information specified in the user profiles (available in 11% of tweets).[1] We first filtered out all tweets from time zones outside North America or that contained location information from other countries in the user profiles (e.g., “Paris, France”). Then, we used the coordinate information of each tweet (if available) to map each tweet to a county. We also used the location information in the users’ profile to determine the county and state of each tweets. When the location information was complete and included both city and state (e.g., “San Diego, California”), we matched the tweet to the relevant county. When the location field was incomplete, i.e., city name, we only matched counties if the name was unambiguous (e.g., “Chicago” was unambiguously Chicago, Illinois, whereas “Urbana” could be in Illinois, Ohio, or Maryland).

Approximately 19% of the tweets could be mapped to U.S. counties (about 155 million tweets), and this percentage is similar to the geo-mapping rate (i.e., 20%) reported in previous studies [30]. We also compared the accuracy of our geo-mapping procedure with that of the Google Maps Geocoding API, a commercial mapping service. A random sample of 20,000 tweets (proportional to the distribution of tweets across the years, and with both location and coordinate information available) was extracted from the data set. When this set of tweets was geo-mapped using the geo-mapping program and the Google Maps Geocoding API,[2] 93% of the mappings were in agreement. Thus, our geolocation method was highly accurate and given the volume of mapped tweets (19% of all tweets), highly adequate given our goals.[3]

Topic Extraction

We used a twofold cross-validation to avoid overfitting, meaning that we split the data into two halves, trained the models on one-half of the data, and tested their performance on the other half (out-of-sample prediction). The results are thus less likely to be based on merely spurious correlations. Rather than conducting separate tests to evaluate whether each individual word is a significant predictor, we used a *topic modeling* procedure. Each county was treated as a document (meaning that the words from all tweets from each county were combined into one single word-by-frequency matrix; see Fig. 1) and associated with a list of topics; each topic refers to a group of semantically related words that co-occurred frequently (e.g., one topic included *storage, gospel, trends,* and *louis*).

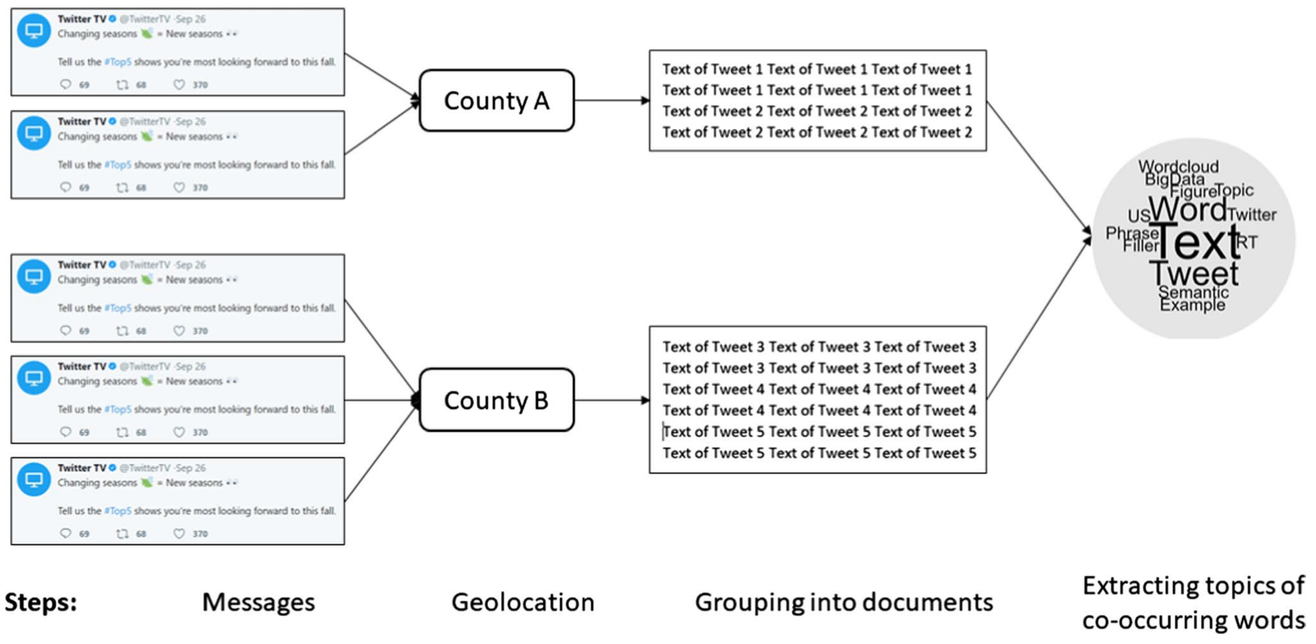
We used the Python package *scikit-learn* to convert a collection of documents (i.e., each county’s tweets) into a frequency matrix of token counts. The matrix of token counts was then analyzed using a well-established algorithm from computer science, Latent Dirichlet Allocation [43]. LDA is a Bayesian mixture model that groups words that often appear together to create topics (see Fig. 1). To illustrate our analytic procedure, we provide more details about using the 2009 model to compute the 2013 ORI of HIV new diagnoses rates based on 2013 tweets. We extracted topics from the 2009 tweets (Fig. 2 illustrates the two semantic features most strongly associated with each outcome of one of the models), and we calculated the presence of each topic in each county with the probability defined as: $p(topic, county) = \sum p(topic|word) \times p(word|county)$, where $p(word|county)$ represents the normalised proportion of words in a county document, and $p(topic|word)$ refers to the probability of the topic given that word.

We next selected the topics that were most predictive of HIV in 2009 using an extremely-randomized-trees model [50]. We chose this method because it retains more topics than other procedures and can thus explain more variance, a choice consistent with both an open-vocabulary approach and our prediction goals. We called this model the 2009 model, and we developed two semantic models for each year, one for the same-year prediction and one for the prediction of different years. Then, we obtained the topics probabilities of the 2013 tweets based on topics of the 2009 model. The topic coefficients of the 2009 model, together with the probabilities of the topics using the 2013 tweets, were entered to calculate the Online Risk Index (ORI) of HIV for 2013 for each county. We correlated these ORI with the 2013 HIV rates reported by the CDC to evaluate the quality of prediction. Fifty topics were extracted that are related to the STI rates, and the probability of each topic per county was estimated based on the relative word frequencies distribution, as described in Schwartz et al. [42].

[1] We compared a set of random tweets with and without location/coordinate information (N = 3,000), which showed remarkably similar vocabulary sizes: yes = 10,911 and no = 11,148, character count (per tweet): yes = 90 and no = 87, and word count (per tweet): yes = 15 and no = 14.

[2] Google Maps Geocoding API is not a free web service (see https://developers.google.com/maps/faq for detailed pricing). Therefore, it is necessary to develop a reliable geo-mapping program for mapping millions of tweets.

[3] About 19% of all tweets can be geo-mapped and left a large part of tweets excluded from the analyses. We can’t assess the model performance of tweets that are with and without location/coordinate information because all HIV/STI new diagnoses rates are reported at the county-level.



Twitter TV @TwitterTV · Sep 26
Changing seasons = New seasons
Tell us the #Top5 shows you're most looking forward to this fall.
69 68 370

Twitter TV @TwitterTV · Sep 26
Changing seasons = New seasons
Tell us the #Top5 shows you're most looking forward to this fall.
69 68 370

Twitter TV @TwitterTV · Sep 26
Changing seasons = New seasons
Tell us the #Top5 shows you're most looking forward to this fall.
69 68 370

Twitter TV @TwitterTV · Sep 26
Changing seasons = New seasons
Tell us the #Top5 shows you're most looking forward to this fall.
69 68 370

Twitter TV @TwitterTV · Sep 26
Changing seasons = New seasons
Tell us the #Top5 shows you're most looking forward to this fall.
69 68 370

Text of Tweet 1 Text of Tweet 1 Text of Tweet 1
Text of Tweet 1 Text of Tweet 1 Text of Tweet 1
Text of Tweet 2 Text of Tweet 2 Text of Tweet 2
Text of Tweet 2 Text of Tweet 2 Text of Tweet 2

Text of Tweet 3 Text of Tweet 3 Text of Tweet 3
Text of Tweet 3 Text of Tweet 3 Text of Tweet 3
Text of Tweet 4 Text of Tweet 4 Text of Tweet 4
Text of Tweet 4 Text of Tweet 4 Text of Tweet 4
Text of Tweet 5 Text of Tweet 5 Text of Tweet 5
Text of Tweet 5 Text of Tweet 5 Text of Tweet 5

Fig. 1 Procedures of geolocation, message grouping, and topic extraction

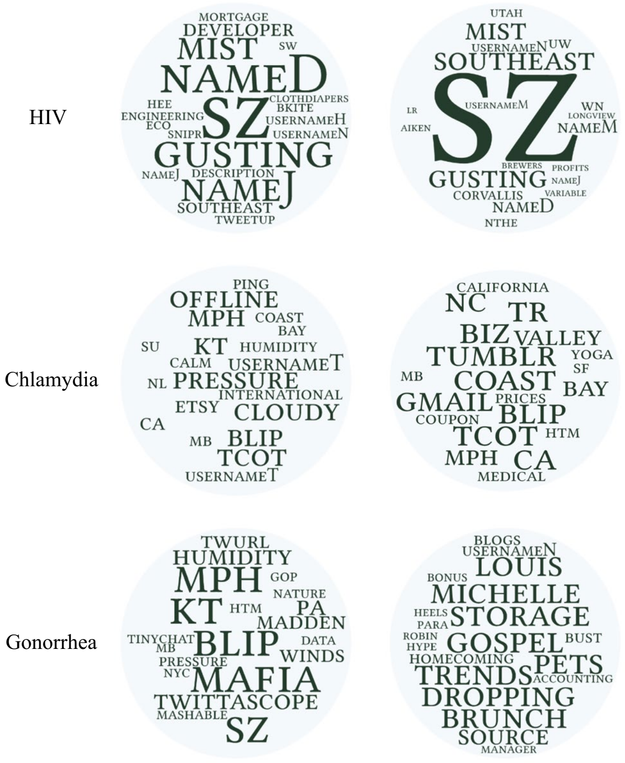


Fig. 2 Top two word clouds (with top 20 words) of the 2009 semantic model to compute the Online Risk Index of STIs in 2013. Terms starting with 'name[CAPITAL LETTER]' or 'username[CAPITAL LETTER]' here are used as substitutes for proper names (usually of celebrities) or usernames to anonymize that data. The capital letter denotes the first letter of the first name. The size of each word indicates the relative weight within a word cloud

Table 2 Number of U.S. counties included in the estimation of ORIs

STIs	2009	2010	2011	2012	2013
Chlamydia	2050	2174	2164	2221	2240
Gonorrhea	1627	1694	1720	1806	1845
HIV	694	674	692	678	687

Statistical Analysis

To reduce the skewness of STI rates, we applied natural log transformations [51]. We trained one model year and used this model to estimate ORIs for all subsequent years. For example, we applied the 2009 HIV topic model (i.e., model of 2009 tweets trained with the HIV rates from 2009) to the 2010 Twitter data to calculate the 2010 ORIs. Next, we input the 2011, 2012, and 2013 Twitter data into the same 2009 model to calculate their respective ORIs. The 2010 model was applied to tweets from 2011, 2012, and 2013 to estimate their ORIs, and so on. To assess the model performance, we examined the correlations between ORIs and actual rates. The analyses were repeated for HIV (N = 674–694 counties; the available number of counties varied across years, see Table 2 for details), chlamydia (N = 2050–2240), and gonorrhea (N = 1627–1845). We also examined three states (California, Florida, and New York) separately to illustrate the usefulness of our models for specific state prediction. All analyses were conducted at the county level.

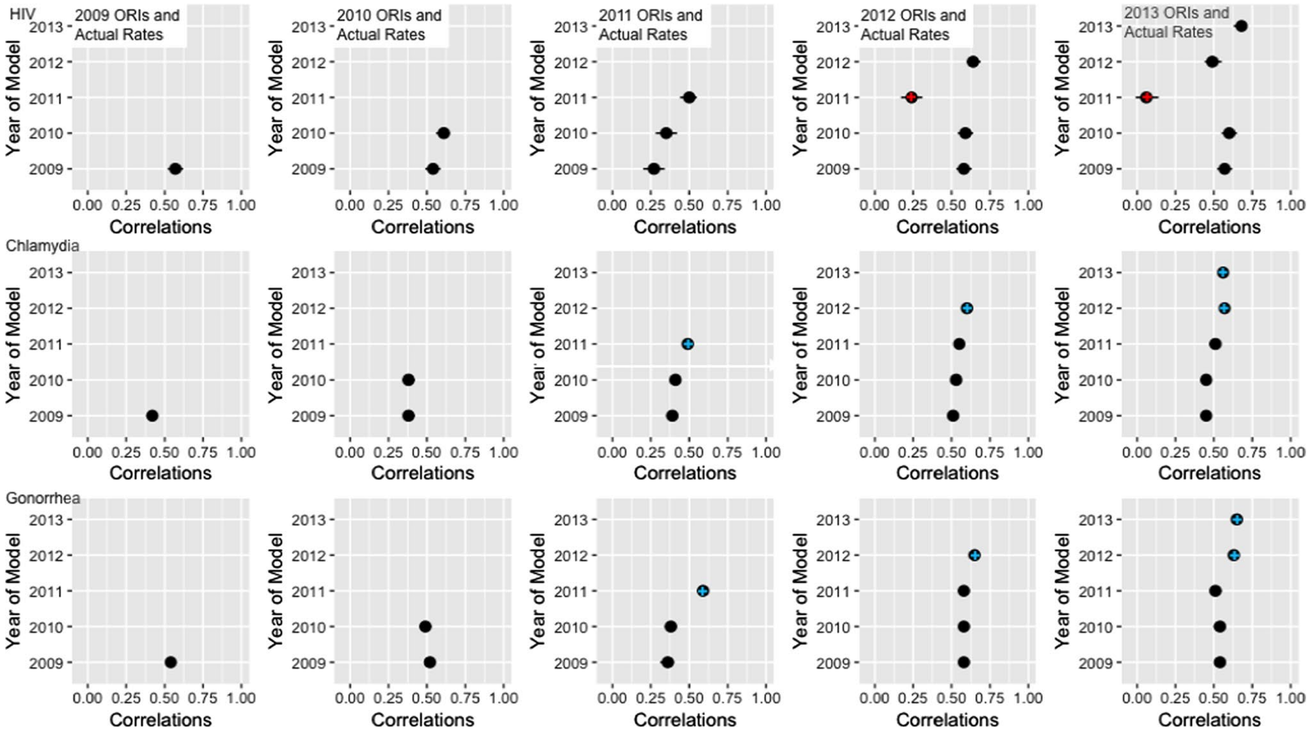


Fig. 3 Forest plots of correlations between the ORIs and actual rates at the county-level in 2009–2013 along with their 95% confidence intervals (error bars). Y-axes indicates which year of model was used to calculate the ORIs and x-axes indicates the correlation levels (Top panels for HIV, middle panels for chlamydia, and bottom panels for gonorrhea). Red dots refer to particularly small correlation coefficients whose confidence intervals do not overlap with the others; blue dots refer to larger correlation coefficients whose confidence intervals do not overlap with the others

Results

Overall Prediction of U.S. Counties

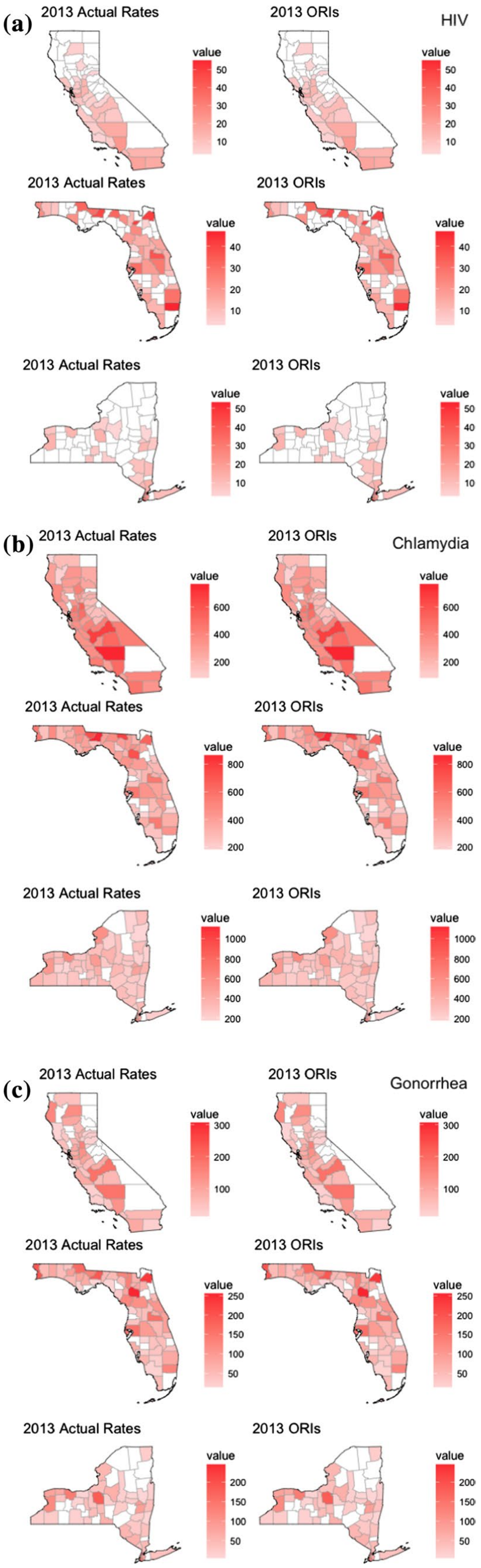
We correlated the ORIs and actual STI rates to assess to what extent social-media-based models can generate useful markers of health status across counties.[4] Figure 3 presents the forest plots of all correlations among the ORIs and actual rates. Most correlations were significant (p values < .0001) and medium (≥ .30) to large (≥ .50) in size [52] (mean correlations of HIV: $r_{\mathrm{m}} = .52$, chlamydia: $r_{\mathrm{m}} = .48$, and gonorrhea: $r_{\mathrm{m}} = .54$). The results indicate that semantic models based on Twitter data can be successfully used to obtain risk indices of HIV, chlamydia, and gonorrhea. This means that social media data from a county, rather than data from only high-risk users, can be used in studying a wide range of STIs. However, the 2011 model showed a few weak correlations, suggesting that models may not perform optimally in all cases: Because our Twitter database contained fewer tweets from 2011 than from other years (see Table 1), it is possible that the small size of the 2011 Twitter data may not be sufficient to identify important semantic features for prediction.

Prediction Across Models Developed in Different Years

Next, Fig. 3 shows a tendency for more recent models to predict the data slightly better than older models, which indicates a *the-later-the-better* pattern. For example, in the analyses of 2013 ORIs, the 95% confidence intervals of 2012 and 2013 models did not overlap with the 2009 and 2010 models (the differences between correlations ranged from .07 to .12). Such a pattern was present in the 2011, 2012, and 2013 ORIs for both chlamydia and gonorrhea (see Fig. 3). However, the *the-later-the-better* pattern was not observed for HIV. The associations between the ORIs and the actual rates were similar with one exception: Again, models trained in 2011 performed significantly worse than models trained in other years. All remaining models performed about equally

[4] We present the rank-based residuals of the actual (non-log-transformed) STI rates and the back-transformed ORIs (see Table in Supplementary Information). The overall residuals showed negligible differences for HIV, gonorrhea, and chlamydia, implying that the semantic models showed no strong biases. Altogether, the semantic models using Twitter language thus provided satisfactory performance in estimating the county-level STI risk.

Fig. 4 **a** Maps of the actual rates (left column) and ORIs (right column) of HIV (top: California, middle: Florida, bottom: New York). Counties shown in white have missing data. **b** Maps of the actual rates (left column) and ORIs (right column) of chlamydia (top: California, middle: Florida, bottom: New York). Counties shown in white have missing data. **c** Maps of the actual rates (left column) and ORIs (right column) of gonorrhea (top: California, middle: Florida, bottom: New York). Counties shown in white have missing data ▶



in predicting risk indices in 2012 and 2013. Taken together, the results provided preliminary evidence that later semantic models can better estimate the ORIs of chlamydia and gonorrhea, but not HIV, and that a large Twitter data set may be necessary for optimal estimation.

Prediction of Specific States

Finally, we singled out California, Florida, and New York to illustrate the prediction of risk indices on a more regional level, rather than nationwide. To visualize the distributions, we plotted the ORIs for 1 year (2013) based on the 2012 model and the actual rates on state maps (Fig. 4). The geographic distributions of the ORIs and the actual rates were consistent and the correlations between these values were moderate to large (California: *r*s = .53–.64, Florida: *r*s = .60–.70, New York: *r*s = .42–.68). Across all years of data, most of the correlations were significant for California and Florida (see Table 3). Model performance for chlamydia and gonorrhea was largely similar across years, indicating that the *the-later-the-better* pattern was not present in California and Florida. As on the national level, the 2011 models performed poorly, especially when base rates were low (HIV) or fewer counties were available (New York). The models of chlamydia and gonorrhea were highly variable for New York: Only 2009 and 2012 chlamydia models and 2010, 2012, and 2013 gonorrhea models performed acceptably. Altogether, these results suggest that the semantic models are valuable on a state-specific level, but not when only a few counties have available data.

Discussion

Our work used Twitter data to obtain online risk indices (ORIs), and the findings showed moderate-to-strong associations between language signals and rates of new HIV, gonorrhea, and chlamydia diagnoses across 5 years of data. The findings demonstrate an overall tendency for models developed in recent years to predict the STI rates slightly better than models developed in earlier years. Overall, we found that text-based social media data can give a useful indication of geographic variation in HIV and STI risk, and such semantic models could also process text data from other social media. The analyses revealed satisfactory

Table 3 Correlations between the 2013 ORIs and the actual rates of new HIV, chlamydia, and gonorrhea diagnoses

STIs	Years of model	States		
		California	Florida	New York
Chlamydia	2013	.58*** [.37, .73]	.57*** [.38, .72]	.26 [.01, .49]
	2012	.56*** [.34, .72]	.60*** [.42, .74]	.42*** [.19, .61]
	2011	.51*** [.29, .69]	.39** [.16, .58]	.13 [-.13, .37]
	2010	.69*** [.51, .81]	.49*** [.28, .66]	.33 [.08, .54]
	2009	.53*** [.31, .70]	.52*** [.32, .68]	.36** [.12, .57]
Gonorrhea	2013	.40** [.13, .61]	.65*** [.48, .78]	.55*** [.33, .72]
	2012	.53*** [.30, .71]	.57*** [.38, .72]	.59*** [.38, .74]
	2011	.52*** [.28, .70]	.51*** [.29, .67]	.33 [.07, .55]
	2010	.55*** [.32, .72]	.34** [.09, .54]	.44*** [.2, .64]
	2009	.42** [.16, .63]	.37** [.13, .57]	.23 [-.04, .47]
HIV	2013	.65*** [.39, .81]	.49*** [.22, .69]	.75*** [.52, .87]
	2012	.64*** [.38, .81]	.70*** [.51, .83]	.68*** [.42, .84]
	2011	.29 [-.06, .58]	.29 [.00, .54]	.16 [-.22, .49]
	2010	.51** [.21, .73]	.55*** [.31, .73]	.64*** [.36, .82]
	2009	.31 [-.03, .59]	.51*** [.25, .70]	.60*** [.30, .79]

Sample sizes of correlation analyses varied among STIs (Chlamydia: *N*s = 54 for California, *N*s = 63 for Florida, *N*s = 59 for New York, Gonorrhea: *N*s = 49 for California, *N*s = 61 for Florida, *N*s = 53 for New York; HIV: *N*s = 33 for California, *N*s = 44 for Florida, *N*s = 29 for New York)

$^{**}p < .01$, $^{***}p < .001$

performance not just nationwide, but also in specific states, offering promising opportunities for integrating big data into local surveillance and prevention work.

In addition to being the first systematic prediction attempt using social media data, our analysis is the first to consider the stability of prediction across years in any disease domain. We improved on previous cross-sectional studies [28–30] by evaluating the prediction performance across models developed for five different years. Two patterns of correlational results were recorded in the present work. The overall strength of the correlations varied minimally across years of input data for HIV: For example, ORIs from models trained in 2010 predicted 2013 HIV ORIs about as well as indices from models trained in 2013 did. Such a pattern was absent only when the Twitter dataset was substantially small, as in our 2011 data.

For chlamydia and gonorrhea, the findings suggest that more recent models are slightly more useful for understanding STIs than older models. Despite this *the-later-the-better* pattern for chlamydia and gonorrhea over three years, the difference in performance was small; the correlations with actual rates differed by about .10. The findings suggest that the prediction via social media data could be designed to minimize effort: Researchers can collect STI data and build a new Twitter-based model every 2–3 years, rather than having to do it every year. Meanwhile, real-time data can still be entered into the model to make time-sensitive estimates with satisfactory performance without waiting for the release of the actual rates to build a model for the current year. In summary, the similarity of prediction with models obtained across years is good news for practical application: A single round of model building can yield a semantic model that predicts well for several years. This allows the implementation of easy-to-use web applications, in which end-users could simply enter text data from a group of counties to obtain ORIs, without developing any machine learning models themselves.

Our findings can provide insights for public health officials to understand and anticipate the STI profile of a region with a relatively low cost and brief time delay. The open-vocabulary approach we selected does not require the identification of location-specific or time-specific words and terms for an area. Twitter-based models may be used to allocate scarce campaign resources into the communities that predictions identify as being at high risk. Furthermore, such ORIs are not only useful in nationwide analyses but also in specific states if enough social media data is available. The advantages of ORIs include that they can be obtained for a widely dispersed population, they do not require expensive setup for the collection of responses, they are based on social media data that can be obtained in real-time, they represent the experiences and communications of a young and diverse population, and they work well over several years. Public health officials, in conjunction with social media analysts, can obtain insights into STIs in areas of interest and add valuable information on decisions of where to allocate resources for further investigation and prevention. In summary, the proposed Twitter-based model can be used to make time-sensitive risk predictions of STIs in communities in an economical and efficient way.

The results presented correlation coefficients and should of course not be the basis for causal conclusions—in many cases, the tweets are likely indicators of social attitudes and conditions in the communities rather than being the cause of disease patterns. Twitter data can thus predict health outcomes for different reasons: The communications might be a signal of distal risk factors which then predict STI risk (e.g., the models may detect counties with higher poverty, which is, in turn, associated with increased risk of HIV), they might reflect proximal risk factors which then predict STI risk (e.g., condomless sex), they might signal infections that have already occurred (e.g., tweets about someone's experiences with a new HIV medication), or they might reflect awareness of STI in the communities (e.g., tweets about news reports or online campaigns that encourage testing). Some of the keywords identified by the topic models also refer to real-world spaces with currently elevated STI rates, such as "southeast" (HIV) or "coast" (chlamydia; Fig. 2). Other words are specific to online spaces, such as usernames of Twitter users or "tweetup" (a meeting that happens on Twitter; Fig. 2). The overall tone of the messages may also convey psychological characteristics that can be predictive of health patterns: For instance, more future-directed tweets have been found to predict lower HIV prevalence [30], and references to positive emotions are associated with lower levels of heart disease mortality [31]. Most likely, the correlations reflect all these aspects.

We recommend that future projects analyze testing data as an additional outcome variable to better disentangle whether different topics are actually associated with higher STI incidence, or merely with higher testing rates. Other work can analyze the message content to identify factors that promote the awareness of HIV campaigns and, potentially, the effects of HIV/STI treatments and care [53]. An examination of the words captured in our semantic models (Fig. 2) nevertheless reveals daily and informal language that seems largely unrelated to testing campaigns or media coverage, thus making it unlikely that the influence of testing campaigns is driving our results.

Further, any causal influences of online social communications on infections may not be reflected in new diagnosis estimates after a certain time interval: For the examination of HIV, the time lag between infection and diagnosis can be several years long. For other infections, the time between infection and testing may range from weeks to months. Prospective data collection of tweets will be necessary to estimate more precisely across which time frame online health campaigns affect STIs in the years to come. Even with the exact nature of the underlying relations unspecified, our findings reveal that tweets can consistently predict the risk of new HIV, chlamydia, and gonorrhea diagnoses even when predicting over a large region and even when the prediction is attempted with models developed in different years. Thus, social media data, specifically tweets, can be a useful information source on which researchers and practitioners can capitalize.

One important caveat is that social-media sites, and the Internet as a whole, experience rapid changes. The approach we used is based on available volume and accessibility of social media data, but is not specific to Twitter. The models of language features will be able to use messages from any text-based sites for which data are available, including, blogs, forums, and chat services. For example, language analyses like ours could also be used for Tumblr posts, Instagram captions, or other text-message clients when such data are available. Specific sites may change, but the centrality of the online world to contemporary life is unlikely to go away.

The present work has other limitations that are worth noting. The proposed ORIs were developed based on a generative probabilistic model which emphasizes explaining the largest variance in most of the cases instead of extreme cases (outliers). Therefore, the ORIs are less likely to estimate counties with extreme values. Future research may develop specific Twitter-based models aimed towards detecting outbreaks of STIs by, for instance, predicting change scores instead of current rates. Furthermore, models for which fewer data were available (e.g., because we had fewer tweets from 2011 than from other years and fewer county-level HIV data are available for New York than for other states) performed poorly compared to other models, highlighting the importance of a large and complete data set for accurate predictions. Additionally, the results from this study are dependent on the quality of the data used to develop the models. Any systematic biases in the surveillance data, such as an underestimation of chlamydia rates in men due to lower testing rates, are not resolved by using social media data. Finally, specific words and phrases identified in the present work might be difficult to interpret within any theoretical framework of STIs due to the diversity of writing styles and conventions used on Twitter.

The results of the present study encourage future research to investigate the causal influences of Twitter as a potential premise of using social media in health campaigns. We present evidence that social media data can be used to estimate risk indices of HIV, gonorrhea, and chlamydia. By analyzing Twitter messages, we can identify geographic regions that are at high risk. Further, the associations between language use and STI rates seem temporally stable and thus useful for predicting health patterns over several years. Whereas official CDC STI data provide precise and reliable estimates after a comparatively long delay in publishing, Twitter data can offer real-time risk indices and can thus be a useful supplement to obtain information quickly. In the future, the ready availability and high temporal resolution of social media data can help researchers, communities, and public health officials to better understand the health status of communities and target timely campaigns in the United States and potentially in other parts of the world.

Acknowledgements This work was funded by a National Institutes of Health grant. We are grateful to Travis Sanchez, Patrick S. Sullivan, and Yisi Liu for their help in data collection.

Funding This study was funded by the National Institutes of Health (Grant Number R56 AI114501 to D. A.).

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Centers for Disease Control and Prevention. *Sexually Transmitted Disease Surveillance 2016*. Atlanta, GA; 2017
2. Centers for Disease Control and Prevention. NCHHSTP AtlasPlus. https://www.cdc.gov/nchhstp/atlas/. Accessed 25 May 2017.
3. Owusu-Edusei K, Chesson HW, Gift TL, et al. The estimated direct medical cost of selected sexually transmitted infections in the United States, 2008. Sex Transm Dis. 2013;40(3):197–201. https://doi.org/10.1097/OLQ.0b013e318285c6d2.
4. Himmelstein DU, Woolhandler S. Public health's falling share of US health spending. Am J Public Health. 2016;106(1):56–7. https://doi.org/10.2105/AJPH.2015.302908.
5. Centers for Disease Control and Prevention. Overview of the CDC FY 2018 budget request. 2017. https://www.cdc.gov/budget/documents/fy2018/fy-2018-cdc-budget-overview.pdf.
6. Garcia-Calleja JM, Jacobson J, Garg R, et al. Has the quality of serosurveillance in low- and middle-income countries improved since the last HIV estimates round in 2007? Status and trends through 2009. Sex Transm Infect. 2010;86(Suppl 2):ii35–ii42. https://doi.org/10.1136/sti.2010.043653.
7. Davis SL, Goedel WC, Emerson J, Guven BS. Punitive laws, key population size estimates, and Global AIDS response progress reports: an ecological study of 154 countries. J Int AIDS Soc. 2017;20(1):21386. https://doi.org/10.7448/IAS.20.1.21386.
8. Sun CJ, Reboussin B, Mann L, Garcia M, Rhodes SD. The HIV risk profiles of Latino sexual minorities and transgender women who use websites and mobile apps designed for social and sexual networking. Heal Educ Behav. 2016;43(1):86–93. https://doi.org/10.1177/1090198115596735.
9. Ayers JW, Althouse BM, Dredze M, Leas EC, Noar SM. News and internet searches about human immunodeficiency virus after Charlie Sheen's disclosure. JAMA Intern Med. 2016;176(4):552. https://doi.org/10.1001/jamainternmed.2016.0003.
10. Aicken CR, Estcourt CS, Johnson AM, Sonnenberg P, Wellings K, Mercer CH. Use of the internet for sexual health among sexually experienced persons aged 16 to 44 years: evidence from a nationally representative survey of the British population. J Med Internet Res. 2016;18(1):e14. https://doi.org/10.2196/jmir.4373.
11. Young SD, Nianogo RA, Chiu CJ, Menacho L, Galea J. Substance use and sexual risk behaviors among Peruvian MSM social media users. AIDS Care. 2016;28(1):112–8. https://doi.org/10.1080/09540121.2015.1069789.
12. Saberi P, Johnson MO. Correlation of Internet use for health care engagement purposes and HIV clinical outcomes among HIV-positive individuals using online social media. J Health Commun. 2015;20(9):1026–32. https://doi.org/10.1080/10810730.2015.1018617.
13. Leite L, Buresh M, Rios N, Conley A, Flys T, Page KR. Cell phone utilization among foreign-born Latinos: a promising tool for dissemination of health and HIV information. J Immigr Minor Heal. 2014;16(4):661–9. https://doi.org/10.1007/s10903-013-9792-x.
14. Blackstock OJ, Cunningham CO, Haughton LJ, Garner RY, Norwood C, Horvath KJ. Higher eHealth literacy is associated with HIV risk behaviors among HIV-infected women who use the internet. J Assoc Nurses AIDS Care. 2016;27(1):102–8. https://doi.org/10.1016/j.jana.2015.09.001.
15. Pennise M, Inscho R, Herpin K, et al. Using smartphone apps in STD interviews to find sexual partners. Public Health Rep. 2015;130(3):245–52. https://doi.org/10.1177/003335491513000311.
16. Lenhart A, Purcell K, Smith A, Zickuhr K. Social media & mobile internet use among teens and young adults. Pew Research Center. http://www.pewinternet.org/2010/02/03/social-media-and-young-adults/. Published 2010.
17. Pew Research Center. Social media fact sheet. http://www.pewinternet.org/fact-sheet/social-media/.
18. Benotsch EG, Kalichman S, Cage M. Men who have met sex partners via the Internet: prevalence, predictors, and implications for HIV prevention. Arch Sex Behav. 2002;31(2):177–83. https://doi.org/10.1023/A:1014739203657.
19. Harfenist E, Cohen A. How opioid addicts are using social media to get clean. The Week. April 30, 2017.
20. Saito S, Howard AA, Chege D, et al. Monitoring quality at scale. AIDS. 2015;29:S129–36. https://doi.org/10.1097/QAD.0000000000000713.
21. Bushman FD, Barton S, Bailey A, et al. Bringing it all together. AIDS. 2013;27(5):835–8. https://doi.org/10.1097/QAD.0b013e32835cb785.
22. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev Med An Int J Devoted to Pract Theory. 2014;63:112–5. https://doi.org/10.1016/j.ypmed.2014.01.024.
23. Khoury MJ, Ioannidis JPA. Medicine. Big data meets public health. Science. 2014;346(6213):1054–5. https://doi.org/10.1126/science.aaa2709.
24. TechAmerica Foundation. Demystifying Big Data: A Practical Guide to Transforming the Business of Government. 2012. https://bigdatawg.nist.gov/_uploadfiles/M0068_v1_3903747095.pdf.
25. Signorini A, Segre AM, Polgreen PM. The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS ONE. 2011;6(5):e19467. https://doi.org/10.1371/journal.pone.0019467.
26. Aslam AA, Tsou M-H, Spitzberg BH, et al. The reliability of tweets as a supplementary method of seasonal influenza surveillance. J Med Internet Res. 2014;16(11):e250. https://doi.org/10.2196/jmir.3532.
27. Santos JC, Matos S. Analysing Twitter and web queries for flu trend prediction. Theor Biol Med Model. 2014;11(Suppl 1):S6. https://doi.org/10.1186/1742-4682-11-S1-S6.
28. Young SD. Behavioral insights on big data: using social media for predicting biomedical outcomes. Trends Microbiol. 2014;22(11):601–2. https://doi.org/10.1016/j.tim.2014.08.004.
29. Ireland ME, Chen Q, Schwartz HA, Ungar LH, Albarracín D. Action tweets linked to reduced county-level HIV prevalence in the United States: online messages and structural determinants. AIDS Behav. 2016;20(6):1256–64. https://doi.org/10.1007/s10461-015-1252-2.
30. Ireland ME, Schwartz HA, Chen Q, Ungar LH, Albarracín D. Future-oriented tweets predict lower county-level HIV prevalence

in the United States. Heal Psychol. 2015;34(Suppl):1252–60. https://doi.org/10.1037/hea0000279.

31. Eichstaedt JC, Schwartz HA, Kern ML, et al. Psychological language on Twitter predicts county-level heart disease mortality. Psychol Sci. 2015;26(2):159–69. https://doi.org/10.1177/0956797614557867.
32. Twitter. Twitter usage. Company Facts. https://about.twitter.com/company. Published 2016.
33. Twitter. Getting started with Twitter. The Basics. https://support.twitter.com/articles/215585. Published 2016. Accessed 18 April 2016.
34. Statista. Social media: daily usage in selected countries as 4th quarter 2015 (fee-based). Social Media & User-Generated Content. http://www.statista.com/statistics/270229/usage-duration-of-social-networks-by-country/. Published 2015. Accessed April 18, 2016.
35. Lenhart A, Smith A, Anderson M, Duggan M, Perrin A. Teens, technology and friendships. Pew Research Center. http://www.pewinternet.org/2015/08/06/teens-technology-and-friendships/. Published 2015. Accessed 21 March 2016.
36. Greenwood S, Perrin A, Duggan M. Social Media Update 2016. 2016. http://www.pewinternet.org/2016/11/11/social-media-update-2016/.
37. Schwartz HA, Eichstaedt JC, Kern ML, et al. Characterizing geographic variation in well-being using tweets. In: Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM). Boston, MA 2013.
38. Schwartz HA, Giorgi S, Sap M, Crutchley P, Eichstaedt JC, Ungar LH. DLATK: Differential language analysis ToolKit. In: Proceedings of the 2017 EMNLP system demonstrations. 2017:55–60.
39. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language use: our words, our selves. Annu Rev Psychol. 2003;54:547–77. https://doi.org/10.1146/annurev.psych.54.101601.145041.
40. Lazer D, Kennedy R, King G, Vespignani A. The parable of google flu: traps in big data analysis. Science (80-). 2014;343(6176):1203–1205. https://doi.org/10.1126/science.1248506.
41. Gouws S, Metzler D, Cai C, Hovy E, Rey M. Contextual bearing on linguistic variation in social media. In: Proceedings of the workshop on languages in social media. 2011, pp. 20–29
42. Schwartz HA, Eichstaedt JC, Kern ML, et al. Personality, gender, and age in the language of social media: the open-vocabulary approach. PLoS ONE. 2013;8(9):e73791. https://doi.org/10.1371/journal.pone.0073791.
43. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. J Mach Learn Res. 2012;3(4–5):993–1022. https://doi.org/10.1162/jmlr.2003.3.4-5.993.
44. Park G, Schwartz HA, Eichstaedt JC, et al. Automatic personality assessment through social media language. J Pers Soc Psychol. 2015;108(6):934–52. https://doi.org/10.1037/pspp0000020.
45. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci. 2013;110(15):5802–5. https://doi.org/10.1073/pnas.1218772110.
46. Karon BP. The clinical interpretation of the Thematic Apperception Test, Rorschach, and other clinical data: a reexamination of statistical versus clinical prediction. Prof Psychol Res Pract. 2000;31(2):230–3.
47. Iacobelli F, Gill AJ, Nowson S, Oberlander J. Large scale personality classification of bloggers. In: Proceedings of the 4th international conference on affective computing and intelligent interaction. New York, NY: Springer 2011:568–577. https://doi.org/10.1007/978-3-642-24571-8_71.
48. Centers for Disease Control and Prevention. National center for health: health indicators warehouse. www.healthindicators.gov. Accessed February 28, 2016.
49. Emory University. Rollins school of public health. AIDSVu. 2016. www.aidsvu.org.
50. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn. 2006;63(1):3–42. https://doi.org/10.1007/s10994-006-6226-1.
51. Howell DC. Statistical methods for psychology. 6th ed. Belmont, CA: Thomson Wadsworth; 2007.
52. Cohen J. A power primer. Psychol Bull. 1992;112(1):155–9. https://doi.org/10.1037//0033-2909.112.1.155.
53. Adrover C, Bodnar T, Huang Z, Telenti A, Salathé M. Identifying adverse effects of HIV drug treatment and associated sentiments using Twitter. JMIR Public Heal Surveill. 2015;1(2):e7. https://doi.org/10.2196/publichealth.4488.