Harnessing digital data and data science to achieve 90–90–90 goals to end the HIV epidemic

Steffanie A. Strathdee, Alicia L. Nobles, and John W. Ayers

Purpose of review
Effective public health interventions depend on timely, accurate surveillance. Harnessing digital data (including internet searches, social media, and online media) and data science is an emerging approach to complement traditional surveillance in public health but has been underutilized in HIV prevention and treatment.

Recent findings
We highlight recent examples that illustrate how social media data can be applied to HIV surveillance and prevention interventions.

Summary
To achieve 90–90–90 goals to end the HIV epidemic, we encourage traditional public health researchers to partner with data scientists to supplement HIV surveillance programs with social media analytics to refine estimates of HIV infections and key populations at risk and to identify subgroups and regions where prevention and treatment efforts need to be bolstered. We also encourage interdisciplinary teams to design interventions to promote HIV prevention and linkage to care by leveraging digital media, such as search engines and social media, that have the potential to reach millions of people instantaneously.

Keywords
digital media, HIV, machine learning, social media, surveillance

DIGITAL DATA AND DATA SCIENCE FOR HIV

Wayne Gretzky allegedly said that he skates to where the puck is *going to be*, not where it has *been*. This quote is commonly referenced by successful industries pointing toward their model to stay one-step-ahead of the market. But can we apply this approach to HIV prevention and treatment to end the epidemic? Herein, we outline how the digital data and data science revolution could impact HIV prevention and treatment efforts.

Effective public health interventions depend on timely, accurate surveillance. Yet almost all existing HIV-related surveillance systems rely on time- and resource-intensive retrospective data that by definition are backward-looking. Databases of AIDS diagnoses suffer from under-, mis-, and delayed-reporting, partially because of the lengthy 10–15-year incubation period for HIV to progress to AIDS. Even in the United States, where we have achieved near-complete data on new HIV diagnoses, these 'new' diagnoses often reflect infections that occurred months or years prior to diagnosis. As a consequence, HIV/AIDS surveillance data is prone to bias and potentially hampers the deployment of timely interventions. Moreover, surveys and in-depth interviews are subject to bias, including nonresponse, missing data, and socially desirable responses, which are problematic given that HIV prevalence is higher among those engaging in highly stigmatized sexual and drug use behaviors.

One approach that holds great promise to propel HIV surveillance and prevention forward is to leverage the power of digital data (including internet searches, social media, and online media) and data science to complement traditional surveillance in public health. Analyses of digital footprints analyses can accurately reflect population-level health and provide more timely data than traditional data

Department of Medicine, Division of Infectious Disease and Global Public Health, UC San Diego, La Jolla, California, USA

Correspondence to Steffanie A. Strathdee, Department of Medicine, Division of Infectious Disease and Global Public Health, UC San Diego, 9500 Gilman Drive, Mail Code 0507, La Jolla CA 92093-0507, USA. Tel.: +1 858 822 1952;. e-mail: sstrathdee@ucsd.edu

Curr Opin HIV AIDS 2019, 14:000–000
DOI:10.1097/COH.0000000000000584

KEY POINTS

- Digital data (including internet searches, social media, and online media) and data science have been underutilized in HIV surveillance and prevention.
- Digital footprint analyses can accurately reflect population-level health and provide more timely data than traditional data collection methods.
- Surveillance of digital media can identify gaps in traditional epidemiologic surveillance methods or service provision to promote uptake of HIV testing and ART.
- Timely analysis of digital media can help fine-tune estimates of HIV-related risk behaviors and monitoring of HIV-related disparities to inform rapid and responsive allocation of HIV prevention and treatment resources.
- By making data freely available to end users, social media data can also help educate and empower HIV-affected communities and encourage HIV testing, uptake of biomedical interventions (e.g., PrEP) and linkage/adherence to HIV care and treatment.

collection methods [1,2]. However, the potential of these data resources and tools in HIV prevention and treatment has yet to be realized. For instance, social media platforms such as Facebook, Instagram, Reddit, and Twitter enable people to share with the public what they are thinking in real-time, often before engaging in specific behaviors. In turn, these publicly available data can be examined using data science principles to detect 'digital biomarkers' such as signals of psychological distress or substance use that are known risk factors related to HIV risk. Applying these data and tools at-scale can yield instantaneous population-level insights to assess behavioral trends that can inform the design and evaluation of HIV prevention programs and interventions. An advantage of surveilling digital media for HIV prevention is that we *directly* observe the data, potentially mitigating the aforementioned biases. Below, we highlight recent examples that illustrate how social media data can be applied to HIV prevention and suggest future directions for harnessing social media to stay one step ahead of HIV and obtain the Joint United Nations Programme on HIV/AIDS (UNAIDS) goals of 90–90–90.

HIV SURVEILLANCE

Publicly available, aggregate data (e.g., Google Trends) and open-source visualization tools (e.g., TobaccoWatcher.org; HealthMap) have been available for over a decade but only recently have they been used to monitor real-time trends in health behaviors, symptoms, and outcomes. These tools have identified infectious disease outbreaks and health trends, including outbreaks of influenza, measles, and food-borne disease as well as trends in smoking and vaping, long before traditional surveillance methods [3–6].

Analysis of social media and corresponding geospatial visualizations can have important implications for HIV surveillance. For example, Young *et al.* [7] found a significant positive relationship between HIV diagnoses tracked through AIDSVU (AIDSVU.org), HIV-related tweets, and aggregated Google search engine queries [8▪▪] and created a near real-time, interactive map based on these digital proxies that tracks diagnoses of HIV in the United States at the national, state, and local level. Studies such as these show the feasibility and predictive potential of using digital data to monitor and evaluate HIV risk behaviors and outcomes.

Surveillance of digital media can also identify gaps in traditional epidemiologic surveillance methods or service provision. For example, Granich *et al.* [9] compared registered users from Hornet [10], a dating app for men having sex with men (MSM) to UNAIDS estimates from 29 countries and found significant discrepancies between official estimates of MSM and the number of individual Hornet users. In particular, they observed 30% more users in 10 of the countries, suggesting that traditional methods for estimating MSM population size may be underestimates that could be enhanced by analyses of social media data. Similarly, a recent study of HIV-related posts on Baidu Tieba, the largest Chinese communication platform (similar to Reddit), found that the number of HIV-related social support requests was approximately three-fold higher than the number of posts providing social support, indicating an appreciable gap in China's HIV service provision [11▪]. Nobles *et al.* [12▪▪] applied natural language processing techniques to Reddit posts to examine information needs of the subreddit 'r/STD' and found that people commonly seek information on risk associated with events of perceived exposure, transmission, symptoms of HIV infection, testing services, test window periods, and interpretation of test results. These examples illustrate how timely analysis of digital media can help fine-tune estimates of HIV-related risk behaviors and monitoring of HIV-related disparities to inform rapid and responsive allocation of HIV prevention and treatment resources. By making data freely available to end users, social media data can also help educate and empower HIV-affected communities and encourage health-promoting behaviors like HIV testing [13].

Chary applied natural language processing techniques to tweets to examine trends in prescription

opioid misuse, a known risk factor for future injection drug users [14], and found that mentions of opioid misuse were closely associated with state-by-state surveillance estimates from the National Surveys on Drug Use and Health [15]. This suggests that social media can be used to provide insights for syndromic toxicosurveillance that could inform HIV prevention efforts among substance users.

Although Twitter is the most common platform for academic research because of the ease of acquiring data, there are a plethora of other social media platforms that cater to specific subpopulations. Instagram and YouTube are popular among young people, partially because of their use of visual imagery. Moovz is the largest exclusively Lesbian, Gay, Bisexual, Transgender social networking site. Reddit and Topix are wide-ranging social media discussion sites that contain communities with HIV-specific discussions or general discussions by at-risk populations such as MSM. MedHelp is a healthcare-focused forum for patients to share information on treatments and prevention, including sites dedicated to discussing HIV. Gay Speak is a forum dedicated to issues facing gay and bisexual young adults. Susan's Place is a forum that houses one of the largest online transgender communities (more than 150 000 posts) in the United States. Platforms such as Craigslist and Reddit, and dating apps, such as Tinder, Bumble, and Grindr, are actively used by people to seek sex partners, although access to the latter is now restricted.

Like any new challenge, applying data science to digital media can pose new challenges and biases, most of which can be anticipated and overcome. The greatest barrier to digital media analytics is access and computational expertise to rapidly analyze large volumes of data across a variety of data types. By partnering with data and computer scientists who are equipped to computationally analyze large volumes of Big Data, public health researchers can collaborate to build automated, scalable, data-driven analyses that can translate into actionable knowledge and strategies for HIV prevention and control. This approach is an advantage over previous HIV research that has primarily relied on selecting small corpuses of data based on keywords and qualitatively analyzing a manageable handful of data. Moreover, because many of the applied analytic strategies are data-driven and largely assumption-free, it is possible to overcome the usual investigator biases in research question selection and measurement. Although storage of large volumes of data has been raised as a concern, accessible metadata repositories and cloud-based services now exist making data management and storage more feasible.

To date, most digital media studies have been limited to media authored in English or Chinese languages [11▪,16]. Exceptions include surveillance of search queries for HIV and sexually transmitted infections (STIs) in Russian conducted using the Yandex search engine [17], and a study of adolescent Twitter users in Botswana that examined HIV risk behaviors posted in English and Setswana [18].

Although social media platforms generally cater to low income, low education, and minority populations [19], who are often more at risk for HIV, some subpopulations such as those in low or middle-income countries, the homeless or people who inject drugs may not access social media as much as others. However, this is beginning to change. In 2014, 58% of people who inject drugs in San Diego, California, reported having routine Internet access [20]. In a more recent study of patients enrolled in an in-patient detoxification clinic in the United States [21], 86% had a mobile phone and almost half had routine daily or weekly internet access on a desktop computer. Over one-third had used internet searches to seek substance abuse treatment. Moreover, the least developed countries are making notable progress toward narrowing the digital divide with the infusion of smartphones [22] and other digital devices.

Not all social media data can be geo-tagged. Only 5% of Twitter users post tweets that include their geo-location; however, those that do provide very precise locations with geo-located HIV-related tweets consistent with geospatial trends from traditional HIV surveillance data [7]. Even when social media cannot be geo-located in real-time, it is sometimes possible to obtain geo-locator data. Beletsky *et al.* [23] developed a method to geo-locate places where study participants in Tijuana, Mexico, lived and engaged in HIV risk behaviors by asking them to point to locations on an interactive Google map.

Data quality can also be an issue, as social media posts are usually anonymous and some posters may deliberately mislead or misrepresent themselves. Social media platforms have also been used to deliberately spread misinformation (e.g., heightening concerns about vaccine safety and efficacy [24]). Public health and HIV prevention researchers should seize the opportunity to leverage the power of the internet for effective health communication, which could include dispelling myths or fears about HIV transmission routes and antiretroviral treatments such as highly active antiretroviral therapy (ART) or pre-exposure prophylaxis (PrEP).

INTERVENTIONS TO PROMOTE HIV PREVENTION AND CONTROL

Digital health interventions are attractive because of their cost-effectiveness and ability to reach millions

of individuals who are otherwise isolated by geography, stigma, or cultural norms. These online venues can be used to recruit cohorts with specific characteristics or to evaluate the efficacy or effectiveness of interventions by monitoring individuals' postings about HIV, risk behaviors, attitudes, and/or intentions [25].

There is an emerging body of literature involving digital media to promote HIV prevention. Several studies have examined the potential for social media interventions to promote HIV testing, linkage to HIV care, or linkage to community-based organizations, especially among MSM [26–28,29,30–32]. In a recent study of MSM using a dating app in China, Wu *et al.* [33▪] developed a scale to identify a subgroup of 'sexual health influencers.' This group had higher rates of testing for HIV and STIs, suggesting that they would be ideal to enlist as 'navigators' to help encourage their social network to engage in safer sex behaviors, HIV testing, uptake of HIV care, and PrEP.

One of the few longitudinal studies to have been conducted examined the effect of recalling, sharing, and participating in visual and text components of a social media intervention on HIV testing among MSM in China [16]. The 1033 men recalled a mean of 2.7 out of six images and shared an average of one image online. Of note is that recalling images/texts or a local contest was associated with facility-based HIV testing.

As adolescents are early adopters of technology, social media platforms are ideally suited for reaching and intervening upon the HIV risk behaviors of young people. In one of the first intervention studies to use social networking sites for HIV prevention, Bull *et al.* [32] conducted a cluster-randomized controlled trial of a Facebook page designed to promote condom use with input from youth. They succeeded in recruiting youth from underrepresented minorities and rural settings and found that exposure to the intervention was significantly associated with greater condom use after 2 months [34]. Future studies should target social media platforms that are most popular among adolescents, such as Instagram, which has more than 1 billion monthly users, whose images can be analyzed using automated image recognition [35]. In addition to prevention, future studies should focus on social media approaches to promote adherence to ART, PrEP, and other interventions.

Real-time monitoring of social media data can identify opportunities where organic responses to current events can be amplified to promote HIV risk reduction or encourage persons testing HIV-positive to seek care. For example, soon after actor Charlie Sheen disclosed his HIV-positive status, members of our group identified record levels of media and online engagement with HIV prevention resources that were the equivalent of seven World AIDS Days [36]. These digital metrics tracked with trends in HIV testing [37]. More importantly, simply publishing this study while the public was still engaged created an echo effect, whereby internet searches for HIV testing increased 12% and HIV self-testing sales increased by 4%. Shortly thereafter, Sheen himself began speaking out on behalf of HIV prevention, citing this work [38].

CONCLUSION

To achieve 90–90–90 goals to end the HIV epidemic, we encourage traditional public health researchers to partner with data scientists to improve uptake of HIV testing, ART, and adherence. Real-time monitoring of HIV surveillance programs with social media analytics can be used to refine estimates of HIV infections and key populations at risk and to identify subgroups and regions where HIV prevention and treatment efforts need to be bolstered. Funding agencies should encourage the training of data scientists in HIV prevention and team grants that have the capability to mine social media data and create new tools to take advantage of them.

Analyses of digital media by cross-disciplinary teams could be transformed into near real-time dashboards [39] such as tobaccowatcher.org that can be easily accessed and utilized by key populations, government agencies, community-based organizations, and thought leaders. To ensure maximum impact and transparency, making analytic tools public and interactive is ideal, so *everyone* with access to the internet has the opportunity to access actionable intelligence to make HIV prevention and control more effective. This would also enable HIV prevention researchers to become more connected to the communities we serve, by rapidly understanding their needs to inform the development of targeted interventions, and in some cases to even amplify organic responses to current events that reduce HIV transmission.

To end the global HIV pandemic, we must skate to where the puck is going to be. Digital media have the potential to monitor and reach populations at risk, including youth, rural communities, mobile populations, and those participating in highly stigmatized risk behaviors. Moreover, digital media platforms can reach both people at risk of acquiring HIV and those that are already HIV infected. The approaches we suggest have the potential to make all HIV prevention and control efforts more evidenced-based, effective (uptake and cost-wise), and enable HIV-affected communities to play an active role in promoting behavior change that they can see in real-time. The digital data and data science revolution are an untapped resource that could

greatly advance 90–90–90 goals to galvanize HIV prevention and control efforts, and ultimately end the HIV pandemic. An advantage of this resource is the potential to reach people on a global scale, not just locally or nationally, to develop innovative interventions that have the potential to reach millions of people at a time, instead of hundreds or thousands.

Acknowledgements

We thank Sharon Park for her assistance in manuscript preparation.

Financial support and sponsorship

Funding for this project was from the California HIV Research Program. S.A.S. also acknowledges support from the National Institute on Drug Abuse (NIDA) through a MERIT award (R37 DA019829) and A.N. acknowledges support from NIDA through a T32 grant (T32 DA023356).

Conflicts of interest

There are no conflicts of interest.

REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:

- ■ of special interest
- ■■ of outstanding interest

1. Ayers JW, Althouse BM, Dredze M. Could behavioral medicine lead the web data revolution? JAMA 2014; 311:1399–1400.
2. Paul MJ, Dredze M. Social monitoring for public health. Synth Lect Inf Concepts, Retrieval, Services 2017; 9:1–183.
3. Brownstein JS, Freifeld CC, Madoff LC. Digital disease detection: harnessing the Web for public health surveillance. N Engl J Med 2009; 360:2153–2155; 2157.
4. Salathe M, Freifeld CC, Mekaru SR, *et al.* Influenza A (H7N9) and the importance of digital epidemiology. N Engl J Med 2013; 369:401–404.
5. Ayers JW, Ribisl KM, Brownstein JS. Tracking the rise in popularity of electronic nicotine delivery systems (electronic cigarettes) using search query surveillance. Am J Prev Med 2011; 40:448–453.
6. Althouse BM, Scarpino SV, Meyers LA, *et al.* Enhancing disease surveillance with novel data streams: challenges and opportunities. EPJ Data Sci 2015; 4:17.
7. Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Prev Med 2014; 63:112–115.
8. ■■ Young SD, Zhang Q. Using search engine big data for predicting new HIV diagnoses. PLoS One 2018; 13:e0199527.

This article shows that Google Trends is a feasible tool to predict new cases of HIV at the state level. The authors discuss the implications of integrating visualization maps and tools based on their models into public health and HIV monitoring and surveillance.

9. Granich R, Gupta S, Garner A, Howell S. Review of UNAIDS national estimates of men who have sex with men, gay dating application users, and 90-90-90 data. bioRxiv 2017; 186163.
10. Hornet Networks L. Hornet. Available at: http://love.hornetapp.com/ (Accessed May 28).
11. ■ Dong Y, Zhou X, Lin Y, *et al.* HIV-related posts from a Chinese internet discussion forum: an exploratory study. PLoS One 2019; 14:e0213066.

This article showed how social media data can augment HIV surveillance data in China. It also showed that the number of HIV-related social support requests was approximately three-fold higher than the number of posts providing social support, indicating an appreciable gap in services.

12. ■■ Nobles AL, Dreisbach CN, Keim-Malpass J, Barnes LE. Is this a STD? Please help!: online information seeking for sexually transmitted diseases on Reddit. Proc Int AAAI Conf Weblogs Soc Media 2018; 2018:660–663.

This seminal report provides several examples of how data from an interactive map of the United States showing the impact of HIV at national, state, and local levels (AIDSVu) can be used to fine-tune the assessment of HIV-related disparities at a community level, educate and empower communities about HIV and its consequences, and better target HIV interventions to reach underserved, vulnerable populations.

13. Valdiserri RO, Sullivan PS. Data visualization promotes sound public health practice: the AIDSvu example. AIDS Educ Prev 2018; 30:26–34.
14. Van Handel MM, Rose CE, Hallisey EJ, *et al.* County-level vulnerability assessment for rapid dissemination of HIV or HCV infections among persons who inject drugs, United States. J Acquir Immune Defic Syndr 2016; 73:323–331.
15. Chary M, Genes N, Giraud-Carrier C, *et al.* Epidemiology from tweets: estimating misuse of prescription opioids in the USA from social media. J Med Toxicol 2017; 13:278–286.
16. Cao B, Saha PT, Leuba SI, *et al.* Recalling, sharing and participating in a social media intervention promoting HIV Testing: a longitudinal analysis of HIV testing among MSM in China. AIDS Behav 2019; 23:1240–1249.
17. Domnich A, Arbuzova EK, Signori A, *et al.* Demand-based web surveillance of sexually transmitted infections in Russia. Int J Public Health 2014; 59:841–849.
18. Cornelius J, Kennedy A, Wesslen R. An examination of twitter data to identify risky sexual practices among youth and young adults in Botswana. Int J Environ Res Public Health 2019; 16:656.
19. Pew Research Center, Greenwood S, Perrin A, Duggan M. Social media update 2016: Facebook usage and engagement is on the rise, while adoption of other platforms holds steady. Washington, DC, 2016 (November 11, 2016).
20. Collins KM, Armenta RF, Cuevas-Mota J, *et al.* Factors associated with patterns of mobile technology use among persons who inject drugs. Subst Abus 2016; 37:606–612.
21. Tofighi B, Leonard N, Greco P, *et al.* Technology use patterns among patients enrolled in inpatient detoxification treatment. J Addict Med 2019; 13:279–286.
22. United Nations Office of the High Representative for the Least Developed Countries LDCaSIDSU-O. World's least developed countries on target to achieve universal and affordable internet by 2020. New York, NY, 2018.
23. Beletsky L, Arredondo J, Werb D, *et al.* Utilization of Google enterprise tools to georeference survey data among hard-to-reach groups: strategic application in international settings. Int J Health Geogr 2016; 15:24.
24. Dredze M, Wood-Doughty Z, Quinn SC, Broniatowski DA. Vaccine opponents' use of Twitter during the 2016 US presidential election: implications for practice and policy. Vaccine 2017; 35:4670–4672.
25. Bauermeister JA, Golinkoff JM, Horvath KJ, *et al.* A multilevel tailored web app-based intervention for linking young men who have sex with men to quality care (get connected): protocol for a randomized controlled trial. JMIR Res Protoc 2018; 7:e10444.
26. Cao B, Gupta S, Wang J, *et al.* Social media interventions to promote HIV testing, linkage, adherence, and retention: systematic review and meta-analysis. J Med Internet Res 2017; 19:e394.
27. Rhodes SD, McCoy TP, Tanner AE, *et al.* Using social media to increase HIV testing among gay and bisexual men, other men who have sex with men, and transgender persons: outcomes from a randomized community trial. Clin Infect Dis 2016; 62:1450–1453.
28. Young SD, Cumberland WG, Nianogo R, *et al.* The HOPE social media intervention for global HIV prevention in Peru: a cluster randomised controlled trial. Lancet HIV 2015; 2:e27–e32.
29. Krueger EA, Chiu CJ, Menacho LA, Young SD. HIV testing among social media-using Peruvian men who have sex with men: correlates and social context. AIDS Care 2016; 28:1301–1305.
30. Cao B, Liu C, Durvasula M, *et al.* Social media engagement and HIV testing among men who have sex with men in China: a nationwide cross-sectional survey. J Med Internet Res 2017; 19:e251.
31. Tang W, Wei C, Cao B, *et al.* Crowdsourcing to expand HIV testing among men who have sex with men in China: a closed cohort stepped wedge cluster randomized controlled trial. PLoS Med 2018; 15:e1002645.
32. Tucker JD, Cao B, Li H, *et al.* Social media interventions to promote HIV testing. Clin Infect Dis 2016; 63:282–283.
33. ■ Wu D, Tang W, Lu H, *et al.* Leading by example: web-based sexual health influencers among men who have sex with men have higher HIV and syphilis testing rates in China. J Med Internet Res 2019; 21:e10171.

This study describes approaches for identifying sexual health influencers using web-based approaches. The authors propose that leveraging these influencers could promote testing for HIV and STDs.

34. Bull SS, Levine DK, Black SR, *et al.* Social media-delivered sexual health intervention: a cluster randomized controlled trial. Am J Prev Med 2012; 43:467–474.
35. Nobles AL, Leas EC, Latkin C, *et al.* #HIV: HIV prevention and treatment advocacy on Instagram. AIDS Behav 2019; Forthcoming.
36. Ayers JW, Althouse BM, Dredze M, *et al.* News and internet searches about human immunodeficiency virus after Charlie Sheen's disclosure. JAMA Intern Med 2016; 176:552–554.
37. Allem JP, Leas EC, Caputi TL, *et al.* The Charlie Sheen effect on rapid in-home human immunodeficiency virus test sales. Prev Sci 2017; 18:541–544.
38. LELOHEX. Charlie Sheen Talks condoms for LELO HEX|Youth is wasted on the young. In: YouTube; 2016.
39. Cohen JE, Ayers JW, Dredze M. Tobacco watcher. Available at: https://tobaccowatcher.globaltobaccocontrol.org/. [Accessed 23 May 2019].