HHS Public Access

Author manuscript

Prev Med. Author manuscript; available in PMC 2016 January 01.

Published in final edited form as:

Prev Med. 2015 January ; 70: 17–18. doi:10.1016/j.ypmed.2014.11.002.

A "big data" approach to HIV Epidemiology and Prevention

Sean D. Young, PhD, MS

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

Abstract

The recent availability of "big data" from social media and mobile technologies provides promise for development of new tools and methods to address the HIV epidemic. This manuscript presents recent work in this growing area of bioinformatics, digital epidemiology, and disease modeling, describes how it can be applied to address HIV prevention, and presents issues that need to be addressed prior to implementing a mobile technology big-data approach to HIV prevention.

Keywords

digital epidemiology; social media; HIV prevention; big data

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

Although HIV remains a tremendous public health challenge after 3 decades of prevention and treatment efforts, the recent availability of "big data" from new technologies provides promise for development of new tools and methods to address the HIV epidemic.

In 2011, it was estimated that more than 1.1 million people were living with HIV/AIDS and 50,000 people were newly diagnosed with HIV in the United States (CDC, 2013). The challenge to combat the spread of HIV is particularly salient among men who have sex with men (MSM), as in 2010, more than half of newly diagnosed HIV cases were among MSM. Traditional public health strategies struggle to reach MSM, leading MSM to be less likely to be test for HIV, access and be retained in care, adhere to treatment, and survive 5 years after diagnosis (Bogart, Wagner, Galvan, & Banks, 2010; CDC, 2002; Hall, Byers, Ling, & Espinoza, 2007). Innovative strategies are needed to provide new tools and better methods of disease surveillance to improve HIV prevention and treatment and reduce the disparities among all populations affected by HIV.

The flood of "big data" from mobile technologies, such as social media, mobile phones, and mobile applications, provides the promise to be able to use these data to develop new HIV monitoring and epidemiology methods, and to provide insights on how to improve HIV interventions and respond to disease outbreaks. Because of the large and increasing use of

The authors declare that there are no conflicts of interest.

mobile technologies among African Americans, Latinos, and Gay populations (Smith, 2010; Young, 2012), analyses of big data from mobile technologies might be particularly helpful in addressing HIV prevention and treatment efforts among these high-risk populations (Young & Jaganath, 2013).

Although there is no clear definition, “big data” refers to datasets that are often characterized by their enormity and complexity (Grant, 2012). These large datasets are available because affordable and easy-to-use technologies have increased the ability for public health researchers to generate large amounts of data (Grant, 2012; Lohr, 2012; Marx, 2013; Murdoch TB & Detsky AS, 2013). For example, the Human Genome Project (HGP), completed in 2003, was an international collaboration to sequence all the base pairs in the human genome. Individual labs were tasked to contribute data from certain areas of the human genome to the HGP database. The combination of these data and the additional combined data have made HGP a classic example of big data (genome.gov, 2014). Big data contain not only relational (structured) data that are conventional in most medical and quantitative datasets, but also unstructured (often free-text) data that can be useful for secondary analyses and qualitative epidemiologic measures (Murdoch TB & Detsky AS, 2013). Unstructured data are important in health research because we can use these free-text data to draw inferences about real-time behaviors and sentiments (Lohr, 2012; Young, Rivers, & Lewis, 2014). For example, social media sites and search engines can be used to collect unstructured posts, messages, searches, updates, and tweets from their users and use these data to inform future public health outbreaks. Influenza researchers have used these unstructured social media data (e.g., from Google searches and Tweets) to predict influenza patterns ahead of the Centers for Disease Control and Prevention (CDC) to strengthen public health preparedness (Broniatowski, Paul, & Dredze, 2013; Ginsberg et al., 2009; Polgreen, Chen, Pennock, Nelson, & Weinstein, 2008).

In fact, the majority of work in this area to date has focused on using big data to respond to influenza outbreaks. For example, Google Flu Trends was designed to tally the number of search terms at any given time that were associated with influenza. The Google Flu Trends algorithm looked at searches for terms such as “influenza” and “early signs of the flu” in order to determine whether these search terms could be used to monitor cases of influenza. Using data from the CDC’s surveillance system (ILINet), a consolidated database of influenza cases reported by the CDC, state and local health departments, and health care providers, studies have found a high correlation (> .9) between Google Flu Trends and ILINet (Cook, Conrad, Fowlkes, & Mohebbi, 2011), suggesting the potential for big data bioinformatics approaches such as Google Flu Trends in monitoring influenza outbreaks. Because of the fairly open access to conversations on Twitter through the “Twitterhose” (Young et al., 2014), Aramaki et al (2011) (Aramaki, Maskawa, & Morita, 2011) applied a similar approach looking at Tweet data (from Twitter) in Japan that included keywords associated with flu-like illness (e.g., cough, fever, and chills). They found that these tweets had up to a .97 correlation with reported influenza cases in Japan.

Although these approaches might apply to a broad number of public health topics, such as influenza, diabetes prevention and management, substance abuse, and sexual health, there has been limited or no work that has been conducted on these topics. Research has been

conducted on this topic around HIV epidemiology and prevention, and analysis of social media big data appears to be feasible for use in that area. After filtering tweets for HIV risk-related keywords and phrases suggesting the occurrence of present or future HIV risk (e.g. sexual behaviors and drug use behaviors), researchers found a high correlation between the geography of these County-level HIV-related tweets and actual CDC reported HIV cases (Young et al., 2014). This study provided further evidence that social media data have the potential to provide a more cost-effective and real-time alternative for HIV remote monitoring and surveillance. Social media data have also aided researchers in HIV prevention efforts, such as HIV interventions and the ability to distribute home-HIV testing kits to those in need. After analyzing free-text posts from an online community focused on HIV prevention, one study found that individuals who posted about HIV prevention and testing, compared to those who posted about other topics, wound up being significantly more likely to request an HIV self-testing kit (Young & Jaganath, 2013). These types of insights based on social media could be valuable in providing health departments with information on how many tests or prevention products might be needed, and determining real-time information on where those health services are about to be requested. More research is needed to refine the methods of using big data in HIV as well as other areas of public health, providing an important and necessary opportunity for HIV and public health researchers.

There has already been criticism about some of the current methods of using technology data for monitoring health outcomes, including the reliability and validity of the data and methods (Lazer, Kennedy, King, & Vespignani, 2014), making it important to highlight issues that need to be addressed prior to using big data from technologies for HIV monitoring. First, there are usability issues with big data approaches as many government agencies, local organizations, and even academic public health departments currently lack the infrastructure to handle big data (Grant, 2012; Murdoch TB & Detsky AS, 2013). Traditional statistical infrastructure is not powerful enough to address the complexity of big data and unstructured data (Grant, 2012; Murdoch TB & Detsky AS, 2013). Instead, collaborations between public health researchers and computer scientists trained in machine learning/data mining are encouraged and perhaps essential to provide the necessary infrastructure for storing and analyzing big data. Second, HIV data need to be released and updated frequently in order to better develop methods of using big data to monitor HIV cases. For example, in the HIV twitter study mentioned above, although tweets were retrieved in real-time during 2012, the most current easily accessible HIV data were from cases in 2009. Therefore, the study could not determine whether social media might be used to monitor ongoing and future HIV cases, but rather could only determine an association between tweets and historical HIV cases (Young et al., 2014). Although there might be limited changes in HIV prevalence from one year to the next, providing access to frequently updated data on HIV cases could help to improve statistical models designed for public health departments to monitor and respond to HIV cases.

This manuscript provides a call for researchers to use technology big data and explore how they can be used to develop new methods of monitoring HIV transmission and other public health concerns. Refinement of these digital epidemiology or bioinformatics methods will help to facilitate the transition from research to practice so that public health organizations

can more readily incorporate these approaches into their epidemiology prevention and monitoring efforts.

Acknowledgements

We wish to thank the National Institute of Mental Health (NIMH) (K01 MH 090884) for funding this work.

References

Aramaki, E.; Maskawa, S.; Morita, M. Twitter Catches the Flu: Detecting Influenza Epidemics Using Twitter. Proceedings of the Conference on Empirical Methods in Natural Language Processing; Stroudsburg, PA, USA. Association for Computational Linguistics; 2011. p. 1568-1576.Retrieved from http://dl.acm.org/citation.cfm?id=2145432.2145600

Bogart LM, Wagner G, Galvan FH, Banks D. Conspiracy beliefs about HIV are related to antiretroviral treatment nonadherence among african american men with HIV. Journal of Acquired Immune Deficiency Syndromes (1999). 2010; 53(5):648–655. doi:10.1097/QAI.0b013e3181c57dbc. [PubMed: 19952767]

Broniatowski DA, Paul MJ, Dredze M. National and Local Influenza Surveillance through Twitter: An Analysis of the 2012-2013 Influenza Epidemic. PLoS ONE. 2013; 8(12):e83672. doi:10.1371/journal.pone.0083672. [PubMed: 24349542]

Unrecognized HIV Infection, Risk Behaviors, and Perceptions of Risk Among Young Black Men Who Have Sex with Men --- Six U.S. Cities, 1994--1998. MMWR. 2002; 51(33):733–736. [PubMed: 12201605]

Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 U.S. dependent areas—2011. HIV Surveillance Supplemental Report. 2013; 18(5) Retrieved from http://www.cdc.gov/hiv/statistics/basics/ataglance.html.

Cook S, Conrad C, Fowlkes AL, Mohebbi MH. Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic. PLoS ONE. 2011; 6(8):e23610. doi:10.1371/journal.pone.0023610. [PubMed: 21886802]

genome.gov. [Retrieved August 18, 2014] All About The Human Genome Project (HGP). 2014. from http://www.genome.gov/10001772

Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, Brilliant L. Detecting influenza epidemics using search engine query data. Nature. 2009; 457(7232):1012–1014. doi:10.1038/nature07634. [PubMed: 19020500]

Grant, B. [Retrieved August 18, 2014] The promise of big data | HSPH News | Harvard School of Public Health. 2012. from http://www.hsph.harvard.edu/news/magazine/spr12-big-data-tb-health-costs/

Hall HI, Byers RH, Ling Q, Espinoza L. Racial/Ethnic and Age Disparities in HIV Prevalence and Disease Progression Among Men Who Have Sex With Men in the United States. American Journal of Public Health. 2007; 97(6):1060–1066. doi:10.2105/AJPH.2006.087551. [PubMed: 17463370]

Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. Science. 2014; 343(6176):1203–1205. doi:10.1126/science.1248506. [PubMed: 24626916]

Lohr S. The age of big data. The New York Times. 2012 Retrieved from http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html?pagewanted=all&_r=0.

Marx V. Biology: The big challenges of big data. Nature. 2013; 498(7453):255–260. doi:10.1038/498255a. [PubMed: 23765498]

Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA. 2013; 309(13):1351–1352. doi:10.1001/jama.2013.393. [PubMed: 23549579]

Polgreen PM, Chen Y, Pennock DM, Nelson FD, Weinstein RA. Using Internet Searches for Influenza Surveillance. Clinical Infectious Diseases. 2008; 47(11):1443–1448. doi:10.1086/593098. [PubMed: 18954267]

Smith, A. Mobile Access 2010. 2010. Retrieved from http://www.pewinternet.org/2010/07/07/mobile-access-2010/

Young SD. Recommended guidelines on using social networking technologies for HIV prevention research. AIDS and Behavior. 2012; 16(7):1743–1745. doi:10.1007/s10461-012-0251-9. [PubMed: 22821067]

Young SD, Jaganath D. Online social networking for HIV education and prevention: a mixed-methods analysis. Sexually Transmitted Diseases. 2013; 40(2):162–167. doi:10.1097/OLQ.0b013e318278bd12. [PubMed: 23324979]

Young SD, Rivers C, Lewis B. Methods of using real-time social media technologies for detection and remote monitoring of HIV outcomes. Preventive Medicine. 2014 doi:10.1016/j.ypmed.2014.01.024.