HHS Public Access

Author manuscript

AIDS. Author manuscript; available in PMC 2024 March 19.

Published in final edited form as:

AIDS. 2021 May 01; 35(Suppl 1): S1–S5. doi:10.1097/QAD.0000000000002888.

Power of Big Data in ending HIV

Bankole Olatosi[1,2], **Sten H. Vermund**[3], **Xiaoming Li**[1,4]

[1.]Big Data Health Science Center, University of South Carolina, Columbia, SC 29208

[2.]Department of Health Services Policy and Management, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208

[3.]School of Public Health, Yale University, New Haven, CT 06510

[4.]Department of Health Promotion, Behavior and Education, Arnold School of Public Health, University of South Carolina, Columbia, SC 29208

Abstract

The papers in this special issue of *AIDS* focus on the application of so-called Big Data sciences as applied to a variety of HIV applied research questions in the sphere of health services and epidemiology. Recent advances in technology means that a critical mass of HIV related health data with actionable intelligence is available for optimizing health outcomes and improving and informing surveillance. Data science will play a key but complementary role in supporting current efforts in prevention, diagnosis, treatment, and response needed to end the HIV epidemic. This collection provides a glimpse of the promise inherent in leveraging the digital age and improved methods in big data science to reimagine HIV treatment and prevention in a digital age.

Big Data science (BDS) in healthcare is rapidly developing due to the increased growth in the volume, variety, velocity, and veracity of health-related data.[1–2] This growth is driven by advanced information and communication technologies (e.g., mobile phones, wearable devices, genomics), affordability of high-performance computing, and the transformative power of modern analytic technologies (e.g., artificial intelligence, machine learning, deep learning). Valuable information, insight and intelligence exist in human immunodeficiency virus (HIV) related data, but it remains to be unlocked efficiently using BDS. The United States National Institutes of Health (NIH) issued its first Strategic Plan for Data Science in May 2018 and suggested that the BDS approach will uniquely advance our understanding of disease prevention, identification, control, and treatment in the coming decades and will be a key to reducing the national and global health disease burden, including HIV.[3] The promise for using BDS to identify and manage high-risk and high-cost patients is well documented.[4]

Studies have identified the complementary role machine learning will play in augmenting the work of healthcare providers,[5] while others suggest predictive modeling using electronic health records (EHR) data[6] will drive precision medicine/public health and improve overall healthcare quality.[7] A preponderance of multimodal and multitudinal data sources

Corresponding author: Bankole Olatosi, Ph.D. Big Data Health Science Center, University of South Carolina, 915 Greene Street, Columbia, SC 29208, olatosi@mailbox.sc.edu.

makes this a possibility. The existence of extensive state level enhanced HIV/AIDS Reporting Systems (e-HARS), Ryan White HIV/AIDS client-level data collected through the Services Report (RSR),[8–9] linkable to EHR and other relevant data sources serve as great opportunities to gather intelligence on patterns of health utilization behavior. This makes focusing on the role big data plays in improving health outcomes for people living with HIV (PLWH) an important part of ending the HIV epidemic.[10]

In response to the increasing availability of large and complex data sets for HIV research, we hope this special issue offers a timely and unique avenue to report new research that apply Big Data (e.g., electronic health records, social media data) and innovative BDS techniques (e.g., machine learning, texting mining, natural language processing, deep learning). BDS can help identify gaps in rare, unseen, and otherwise undiscovered biomedical, behavioral, and social determinants that shed light on HIV acquisition, transmission, the development of comorbidities, and long-term viral load control across the HIV treatment continuum.

The application of BDS in HIV is limited due to the structural and methodological challenges associated with data acquisition, analysis, and interpretability. The challenges include issues around patient privacy, ethical use of data, acquisition of data, data use agreements, data sharing, and repurposing and misuse of big data.[11–12] Benefits include the rich actionable insights BDS provides which range from improved clinical decision support, risk identification, disease prediction and clinical care.[2,3,6] BDS also allows individual level prediction and identification of unique patient clusters compared to traditional statistical methods. BDS techniques like machine learning can also better address data complexities associated with the volume, variety, velocity, and veracity of linked HIV data to improve care, anticipate the needs of PLWH and their providers, achieve cost saving, and improve precision medicine/health services for PLWH.

As seen elsewhere in healthcare, the HIV research field has been slow to embrace and leverage the “sea change” in growing linkable healthcare data for improvement.[13–19] The encounter between generating a scientific response to rapidly growing HIV data sources applying both traditional statistics (supervised learning) and machine learning (unsupervised learning), creates a unique opportunity that we must collectively and urgently adopt to gain new insights. While BDS is taken for granted in the “-omics” fields, it is underexploited in public health and clinical outcomes research.

The field of HIV has a tradition of successful research focused on understanding behaviors associated with surveillance, healthcare utilization, and prevention.[20] Perspectives from BDS can only help to add more value, generate actionable insights, and focus on precision health for all PLWH. While the HIV research and stakeholder communities recognize the potential benefits inherent in the application of BDS to large, linked data sources, some doubt its added value. Others remain concerned about privacy and equity issues and some have concerns about interpretability and clinical relevance. To overcome these concerns, we need evidence from a strong group of BDS HIV researchers committed to help document the value and policy impact of this evolution in research strategy as we work towards ending the epidemic using larger databases for informed feedback. This collection of articles in *AIDS*, with a focus on the applicability of BDS to HIV related data, offers a glimpse into the

leading edge of BDS in HIV that we hope will help engage the HIV research community to see future value in leveraging BDS.

One set of articles in this supplement leverages traditional health department data consisting of different combinations of reported HIV surveillance data, electronic health records and other patient level data sources to answer unique questions. Wang et al.[21] employed a novel analytic approach to develop a Learning Framework of Risk Stratification for HIV (ALERT-HIV) in predicting adolescent HIV risk behaviors (multiple sexual partners and no condom use) using comprehensive longitudinal data from an implementation science research in the Bahamas. Machine learning techniques (support vector machine [SVM] and random forests [RF]) were applied to leverage comprehensive longitudinal data for robust HIV risk behavior prediction. This study provides a good example on how BDS can be used to inform precision HIV behavioral prevention so it can move beyond universal interventions to those tailored for high-risk individuals.

Similarly, Xiang and colleagues[22] developed advanced graph-based deep learning models ("Graph Attention Networks" or GATs) to predict HIV infection among young men who have sex with men (YMSM) aged 16–29 years old from two urban cities (Houston and Chicago) between 2014–2016. Further, integration approaches were used to combine both heterogeneous networks based on multi-graph GAT methods. The authors found that the graph-based GAT models considerably improved the prediction of HIV infection, which largely benefited from its capability of identifying influential neighbors within the social network formed by multiple relations comprised of peers, friends, sex partners and a venue co-attendance network, as well as individual-level sociodemographic and sexual behavioral factors. Such novel methods provide a comprehensive and interpretable modeling framework that may lead to new approaches for HIV prevention and disease intervention.

From a treatment perspective, missed opportunities for HIV testing holds significance for ending the epidemic, since early diagnosis is important to the HIV treatment cascade and continuum. Weissman and colleagues[23] used Big Data and machine learning techniques to identify predictors of missed opportunities for HIV testing among PLWH in South Carolina who visited any health care facilities within eight years before HIV diagnosis (for late presenters) or within three years before HIV diagnosis (for non-late presenters). The authors found that prediction models using machine learning techniques can identify predictors of "missed opportunities" for HIV diagnosis. Their study findings hold promise for improving more precise targeting of HIV testing/prevention efforts to improve early diagnosis.

Chen and colleagues[24] applied a machine learning modeling framework to predict delayed linkage to care in patients newly diagnosed with HIV in Mecklenburg County, NC using deidentified surveillance data. The authors also aggregated linkage to care by zip codes to identify high-risk communities within the county. Their findings provide personalized recommendations for individual patients to better understand their own care continuum. The results also provide guidance for public health teams to identify patient clusters at high-risk for delayed HIV care. The methodology framework and insights can provide a more comprehensive understanding of challenges in HIV linkage to care in NC and similar regions with HIV epidemic and challenge of delayed linkage to care. Delayed linkage to

care holds implications for prevention and transmission of HIV for at-risk communities and remains a centerpiece of the HIV treatment continuum.

Olatosi and colleagues[25] applied machine learning techniques to classify the HIV medical care status (in-care vs not-in-care) for PLWH in SC. The authors compared multiple classification algorithms such as deep neural networks, automated neural networks, decision trees and regression and compared models by examining model classification performance (future case prediction, hidden input selection and complexity optimization) using standard machine learning measures and receiver operating curves (ROC). The authors concluded algorithmic applications such as Bayesian network, neural networks and other machine learning techniques hold significant promise for predicting future states of PLWH HIV care status. Their findings also highlight the benefits BDS adds to traditional statistical methods, and more precisely helps predict individuals most at-risk for dropping out of care in the future. Retention in HIV medical care has long been recognized as an important factor for ending the HIV epidemic.[26–27] The costs of finding and reengaging PLWH back into care makes improving retention important.

Yang and colleagues[28] used two machine learning approaches (LASSO regression and classification and regression tree [CART] analysis) to understand predictors of comorbidity among PLWH in SC. Thirty-five risk predictors were used to predict the severity of comorbidity based on a standard Charlson Comorbidity Index (CCI). The authors concluded that the machine learning methods could help identify the most important predictors of future comorbidity among PLWH with high accuracy. Results may enhance the understanding of comorbidity and provide the data-based evidence for future care management of PLWH. This is of significance for HIV as PLWH age and live longer. The ability to use existing Big Data from the EHR will help plan for the management of an aging PLWH population.

Location plays an important role in access to healthcare. Social determinants of health are directly correlated with individual residence. Using the enhanced HIV/AIDS reporting system combined with publicly available data sources, Zeng et al.[29] examined the geospatial variations in retention in care among PLWH in SC as well as the social and environmental predictors of such geospatial variation based on a sociological framework of health using the LASSO regression and random forest analysis. The study showed that both models demonstrated good predictive accuracy and could identify important contextual predictors of county-level retention-in-care status such as poverty proportion, education levels, proportion of health insurance coverage, and unemployment rates. Their findings call for structural level intervention that improves HIV treatment and care for PLWH and highlight the importance of location of patient residence.

Another set of submissions provided evidence for conducting BDS HIV research using social media data. Studies show social media data are valuable for identifying behavioral insights and predicting biomedical outcomes.[30–34] Today, “digital epidemiologists” often complement traditional surveillance and health-related research by adding new insights using BDS and Big Data.[35] Cheng and colleagues[36] applied an interactive deep learning approach to identify HIV related digital social influencers using Twitter data. Out of a

random 1% sample of 1.15 million Twitter users' data from March 2018 to March 2020, the authors extracted tweets from 1,099 Twitter users who had mentioned "HIV" or "AIDS" and identified two Twitter users to be "online HIV influencers" using a graph neural network model. Their efforts demonstrate the viability of identifying influencers based on conversation topics and engagement. Results suggest that iterative deep learning models can be used to automatically identify new and changing key HIV-related influencers across online big data (e.g., hundreds of millions of social media posts per day) to help promote HIV prevention campaigns to affected communities. Due to the transient nature of the digital age, key HIV-influencers will often change, but iterative deep learning models hold promise for identifying such changes and enhancing the real-time promotion of HIV prevention/treatment campaigns, thereby offering enhanced health promotion and outreach approaches based on BDS.

Li and colleagues[37] explored the feasibility of building a social media-based HIV risk behavior index (SRB) at the county level for informing HIV surveillance and prevention using Twitter data. The authors extracted and analyzed 450,000 HIV risk behavior related geotagged tweets from 250,000 twitter users, developed the SRB based on the content of their tweets, and correlated the SRB with county-level HIV incidence data from AIDSvu. This innovative research demonstrated that it is feasible to build a social media based SRB at the county level for informing HIV epidemic surveillance and prevention. The research also highlighted that geolocation is an important factor that needs to be considered in analyzing county level HIV risk behaviors to reveal spatial heterogeneity.

Due to the growth and explosion of numerous social media platforms, it is hard to reach key population groups like men who have sex with men (MSM)/young MSM (YMSM), persons who inject drugs (PWID), and commercial sex workers. As a result, digital strategies for HIV testing and treatment are lagging behind for key populations.[38–39] Chan and colleagues[40] examined associations among social media (Twitter) postings, in-person conversations about HIV issues, HIV prevention and testing, and MSM norms, indexed by estimated county-level MSM rates (per 1,000 adult men). This study provided moderate to very strong evidence that messages on social media can influence individual communicative behaviors and HIV prevention and testing. The results indicate that the presence of higher proportions of MSM in a county provide a leverage point for social media to foster dialogues about health and HIV. The county MSM norms also engendered the necessary level of county-level messaging about HIV issues on social media and thus indirectly facilitated prevention and testing.

While this *AIDS* special issue provides exciting evidence of the opportunities for using Big Data in addressing critical issues in HIV prevention, treatment, and care, its authors also suggest a host of remaining methodological challenges. First, data restrictions limit the size and quality of the data available for analyses which impacts predictive accuracy.[41] Second, due to episodic nature of most healthcare data, temporality becomes a challenging issue to handle during data analyses. Third, interpretability in unsupervised learning approaches is difficult particularly for providers and stakeholders used to traditional statistical methods. However, there are growing options for using traditional statistical methods (e.g., principal component analyses/contingency tables) with unsupervised learning, engaging domain

experts to interpret findings. Third, patient safety, ethics, and confidentiality and privacy issues remain, and efforts must be made to balance the risks and benefits of using Big Data to improve health.[42]

While these challenges limited to generalizability, and sometimes validity to some extent, for some findings, they also provide us with opportunities to improve Big Data research in HIV. Consistent with milestones we have achieved thus far with HIV/AIDS by using research to overcome stigma and fear, improve ART and HIV testing, and understanding the unique sociodemographic, social, and economic drivers for different PLWH, we have can achieve even more using BDS. Given the advances in mobile technologies, wearables, electronic health records, Internet of Things, and social media, it is increasingly necessary to apply BDS to multimodal sources of HIV data.[43–45] Focus and effort needs to be placed on connecting more data sources to improve predictive modeling accuracy, gain new actionable insights and intelligence, and leave no PLWH behind in the efforts to end the epidemic. This diverse collection of applied BDS represents a glimpse of what researchers can do for the HIV field. We sincerely hope to inspire interest towards the application of BDS and development of novel methods.

Acknowledgement

The work by Bankole Olatosi and Xiaoming Li was in part supported by National Institutes of Health grant # R01AI127203 and the University of South Carolina Excellence Initiatives.

References

1. Murdoch TB, Detsky AS. The inevitable application of big data to health care. JAMA 2013; 309(13):1351–1352. [PubMed: 23549579]
2. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. Health Inf Sci Syst 2014; 2(1): 1–10. [PubMed: 25825665]
3. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. Health Aff 2014; 33(7):1123–1131
4. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. New Engl J Med 2019; 380(14): 1347–1358. [PubMed: 30943338]
5. Adler-Milstein J, DesRoches CM, Kralovec P, Foster G, Worzala C, Charles D, et al. Electronic health record adoption In US hospitals: progress continues, but challenges persist. Health Aff 2015; 34: 2174–2180 (2015).
6. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digital Medicine 2018; 1(1): 1–10. [PubMed: 31304287]
7. CDC HIV Surveillance reports - State enhanced HIV/AIDS Reporting Systems. Available at: https://www.cdc.gov/hiv/library/reports/hiv-surveillance.html (Accessed 3/12/21).
8. Zhu J, Fanning M, Sheehan L, Morrissey KG, Legum S, Hermansen S. Methodology for linking Ryan White HIV/AIDS Program Services Report (RSR) client level data over multiple years. PloS one 2020; 15(8): e0237635.
9. Ryan White Services Report. Available at: https://hab.hrsa.gov/data (Accessed 3/12/21).
10. Yu KH, Beam AL, Kohane IS. (2018). Artificial intelligence in healthcare. Nat Biomed Eng 2018; 2(10): 719–731.
11. Bustreo F, Tanner M. (2020). How do we reimagine health in a digital age? Bull World Health Organ 2020; 98(4): 232. [PubMed: 32284642]
12. Paul AK, Schaefer M. Safeguards for the use of artificial intelligence and machine learning in global health. Bull World Health Organ 2020; 98(4): 282. [PubMed: 32284653]

13. Mu Y, Kodidela S, Wang Y, Kumar S, Cory TJ. The dawn of precision medicine in HIV: state of the art of pharmacotherapy. Expert Opin Pharmacother 2018; 19(14): 1581–1595. [PubMed: 30234392]
14. Cusato J, Allegra S, Nicolò AD, Calcagno A, D'Avolio A. Precision medicine for HIV: where are we?. Pharmacogenomics J 2018; 19(2): 145–165.
15. Olatosi B, Zhang J, Weissman S, Hu J, Haider MR, Li X. Using big data analytics to improve HIV medical care utilisation in South Carolina: A study protocol. BMJ open 2019; 9(7): e027688.
16. Young SD. A "big data" approach to HIV epidemiology and prevention. Prev Med 2015; 70: 17–18. [PubMed: 25449693]
17. van Heerden A, Young S. Use of social media big data as a novel HIV surveillance tool in South Africa. Plos one 2020; 15(10): e0239304.
18. Strathdee SA, Nobles AL, Ayers JW. Harnessing digital data and data science to achieve 90–90–90 goals to end the HIV epidemic. Curr Opin in HIV and AIDS 2019; 14(6): 481–485. [PubMed: 31449089]
19. Liang C, Qiao S, Olatosi B, Lyu T, Li X. Emergence and Evolution of Big Data Analytics in HIV Research: Bibliometric Analysis of Federally Sponsored Studies 2000–2019. medRxiv 2021: Posted January 13 2021. doi: 10.1101/2021.01.11.212496242021.01.13).
20. Geng E, Hargreaves J, Peterson M, Baral S. Implementation research to advance the global HIV response: introduction to the JAIDS supplement. J Acquir Immune Defic Syndr 2019;82: S173–S175. [PubMed: 31764251]
21. Wang B, Liu F, Deveaux L, Gosh S, Ash A, Li X, et al. Adolescent HIV-related behavioral prediction using machine learning: a foundation for precision HIV prevention. AIDS. In press.
22. Yang X, Fujimoto K, Schneider J, Li F, Zhi D, Tao C. A data science approach to predict HIV infection among young MSM: Identifying influential neighbors within social and venue co-attendance networks. AIDS. In press.
23. Weissman S, Zhang J, Chen S, Olatosi B, Li X. Big Data and Machine learning to predict missed opportunities for HIV diagnosis in South Carolina. AIDS. In press.
24. Chen S, Owolabi Y, Dulin M, Robinson P, Witt B, Samoff E. Applying a machine learning modeling framework to predict delayed linkage to care in patients newly diagnosed with HIV in Mecklenburg County, NC. AIDS. In press.
25. Olatosi B, Sun X, Zhang J, Liang C, Li X. Application of machine learning techniques in classification of HIV medical care status for people living with HIV in South Carolina. AIDS. In press.
26. Mugavero MJ, Amico KR, Horn T, Thompson MA. The state of engagement in HIV care in the United States: from cascade to continuum to control. Clin Infect Dis 2013; 57:1164–1171. [PubMed: 23797289]
27. Koester KA, Johnson MO, Wood T, Fredericksen R, Neilands TB, Sauceda J, et al. The influence of the'good'patient ideal on engagement in HIV care. Plos one 2019; 14(3): e0214636.
28. Yang X, Zhang J, Chen S, Weissman S, Olatosi B, Li X. Machine learning approaches to understanding predictors of comorbidity among people living with HIV in electronic health record data. AIDS. In press.
29. Zeng C, Zhang J, Sun X, Li Z, Weissman S, Olatosi B, et al. Contextual factors with county-level retention in care status among people living with HIV in South Carolina from 2005 to 2016. AIDS. In press.
30. Young SD. Behavioral insights on big data: using social media for predicting biomedical outcomes. Trends in Microbiology 2014; 22(11): 601–602. [PubMed: 25438614]
31. 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. [PubMed: 24513169]
32. Salathé M. Digital epidemiology: what is it, and where is it going? Life Sci Soc Soc Policy 2018; 14(1): 1–5.
33. Young SD, Yu W, Wang W. Toward automating HIV identification: machine learning for rapid identification of HIV-related social media data. J Acquir Immune Defic Syndr (1999) 2017; 74: (Suppl 2), S128.

34. Park HA, Jung H, On J, Park SK, Kang H. Digital epidemiology: use of digital data collected for non-epidemiological purposes in epidemiological studies. Healthc Inform Res 2018; 24(4): 253. [PubMed: 30443413]
35. Tarkoma S, Alghnam S, Howell MD. Fighting pandemics with digital epidemiology. EClinicalMedicine. Published August 25 2020 DOI:10.1016/j.eclinm.2020.100512.
36. Zheng C, Wang W, Young S. Identifying HIV-related digital social influencers using an iterative deep learning approach. AIDS. In press
37. Li Z, Qiao S, Jiang Y, Li X. Building a social media-based HIV risk behavior index: A feasibility study. AIDS. In press.
38. Campbell CK, Lippman SA, Moss N, Lightfoot M. Strategies to increase HIV testing among MSM: a synthesis of the literature. AIDS Behav 2018; 22(8): 2387–2412. [PubMed: 29550941]
39. Knight V, Wand H, Gray J, Keen P, McNulty A, Guy R. Implementation and Operational Research: Convenient HIV testing service models are attracting previously untested gay and bisexual men: a cross sectional study. J Acquir Immune Defic Syndr. 2015;69(5):e147–55. doi:10.1097/QAI.0000000000000688. [PubMed: 25970653]
40. Chan MS, Morales A, Zlotorzynska M, Sullivan P, Sanches T, Zhai C, et al. Social media postings about HIV and HIV testing and PrEP use among MSM in the United States. AIDS. In press.
41. Paul AK, Schaefer M. Safeguards for the use of artificial intelligence and machine learning in global health. Bull World Health Organ 2020; 98(4): 282. [PubMed: 32284653]
42. Smith MJ, Axler R, Bean S, Rudzicz F, Shaw J. Four equity considerations for the use of artificial intelligence in public health. Bull World Health Organ 2020; 98(4): 290. [PubMed: 32284656]
43. Vermund SH. Use of big data to identify risk of adverse HIV outcomes. Lancet HIV 2019;6(8):e488–e489. [PubMed: 31303556]
44. Rana AI, Mugavero MJ. How Big Data Science Can Improve Linkage and Retention in Care. Infect Dis Clin North Am 2019;33(3):807–815. [PubMed: 31395146]
45. Young SD, Zhang Q. Using search engine big data for predicting new HIV diagnoses. PLoS One 2018;13(7):e0199527.