Machine Learning and Clinical Informatics for Improving HIV Care Continuum Outcomes

Jessica P. Ridgway[1] · Alice Lee[1] · Samantha Devlin[1] · Jared Kerman[1] · Anoop Mayampurath[2]

Abstract

Purpose of Review This manuscript reviews the use of electronic medical record (EMR) data for HIV care and research along the HIV care continuum with a specific focus on machine learning methods and clinical informatics interventions.

Recent Findings EMR-based clinical decision support tools and electronic alerts have been effectively utilized to improve HIV care continuum outcomes. Accurate EMR-based machine learning models have been developed to predict HIV diagnosis, retention in care, and viral suppression. Natural language processing (NLP) of clinical notes and data sharing between healthcare systems and public health agencies can enhance models for identifying people living with HIV who are undiagnosed or in need of relinkage to care. Challenges related to using these technologies include inconsistent EMR documentation, alert fatigue, and the potential for bias.

Summary Clinical informatics and machine learning models are promising tools for improving HIV care continuum outcomes. Future research should focus on methods for combining EMR data with additional data sources (e.g., social media, geospatial data) and studying how to effectively implement predictive models for HIV care into clinical practice.

Keywords Machine learning · EMR · Clinical informatics · HIV

Introduction

Electronic medical record (EMR) adoption has expanded rapidly in the USA in the past decade. As of 2017, 86% of office-based physicians and 96% of US hospitals had adopted advanced EMR systems [1]. Because EMRs are now utilized by the vast majority of HIV medical providers, electronic data captured in EMRs can greatly enhance understanding of HIV epidemiology. Recent advances in artificial intelligence and machine learning methods allow for detection of complex relationships within EMR data. Beyond elucidating patterns in HIV care, EMRs can also be utilized to enact interventions for improving patient health. In providing care for patients, medical providers spend a significant amount of time interacting with the EMR. Clinical informatics tools embedded within the EMR can give relevant information to providers at the point of care. For example, clinical decision support tools can assist providers with identifying patients at risk for HIV or people living with HIV who are in need of relinkage to care. Machine learning algorithms utilizing EMR data can accurately predict potential future events, such as risk for virologic failure, and this information can be shown to providers to allow them to intervene to improve outcomes for patients in real time.

In this review, we discuss the use of EMR data for HIV-related care and research along the HIV care continuum. We specifically focus on the use of machine learning methods applied to EMR data as well as clinical informatics interventions to improve care continuum outcomes. We also discuss challenges in using EMR data and machine learning for HIV research as well as promising future directions for harnessing these technologies to enhance knowledge and improve quality of care for people living with HIV (PLWH).

HIV Diagnosis

To identify PLWH who are as of yet undiagnosed with HIV, EMR data have been utilized for targeted HIV testing programs. Ahlstrom et al. used machine learning algorithms to

---

This article is part of the Topical Collection on *eHealth and HIV*

✉ Jessica P. Ridgway
  Jessica.ridgway@uchospitals.edu

1 Department of Medicine, University of Chicago, 5841 S Maryland Ave, MC 5065, Chicago, IL 60637, USA

2 Department of Pediatrics, University of Chicago, Chicago, IL, USA

create models predicting HIV status within Danish EMR registries [2]. They found that models utilizing past medical history data within the EMR had higher accuracy for identifying undiagnosed PLWH than models only utilizing demographics and history of sexually transmitted infections. In addition to data documented within structured EMR fields (e.g., "past medical history," "problem list," medications, laboratory values), natural language processing (NLP) of unstructured text of clinical notes in the EMR may be able to detect nuanced risk factors for HIV acquisition. Indeed, Feller et al. found that an algorithm utilizing both structured fields and NLP of unstructured clinical notes to predict risk for HIV acquisition was more accurate than an algorithm using structured EMR data alone [3]. Machine learning models for identifying undiagnosed PLWH can be calibrated to different thresholds of sensitivity and specificity depending on a healthcare system's resources for HIV testing.

Beyond targeted HIV screening, EMRs have also been used to facilitate universal HIV screening. To improve rates of HIV diagnosis among PLWH, the Centers for Disease Control and Prevention and United States Preventive Task Force recommend that all patients be screened for HIV [4, 5]. Despite these recommendations, HIV screening rates in healthcare settings remain low [6]. EMR-based clinical decision support (CDS) tools that prompt providers to order HIV screening have been successfully utilized to improve HIV screening rates in a variety of settings including primary care practices and emergency departments [7, 8]. For example, Lin et al. utilized an EMR-driven clinical decision support tool that linked HIV screening with other routine blood tests in the emergency department to increase monthly HIV screening from an average of 7 HIV screens per month to an average of 550 HIV screens per month [7].

Retention in Care

In addition to aiding diagnosis of HIV among PLWH, EMR data have also been utilized to facilitate relinkage to care for PLWH not engaged in medical care. Ridgway et al. developed an EMR algorithm to identify PLWH not engaged in care who presented to the emergency department or were hospitalized [9]. The algorithm included laboratory data, billing diagnoses, past medical history, problem list, and medications. At their institution, an HIV care navigator utilized the EMR algorithm to identify PLWH in need of relinkage to care. In the first year of use, the algorithm facilitated relinkage of two-thirds of out-of-care patients. Other healthcare systems have used EMRs to coordinate supportive care services for PLWH and improve communication between case managers and other supportive service providers; this intervention was associated with significantly improved retention in care [10].

While the above examples highlight the use of EMR data for relinkage within individual healthcare systems, data sharing among different healthcare systems and/or with the public health department can provide further support for relinkage to care. One of the first examples of such a data sharing approach was the Louisiana Health Information Exchange (LaPHIE), a bi-directional data exchange platform that linked HIV surveillance data from the Louisiana Office of Public Health with patient-level EMR data from Louisiana State University Health Care Services Division (LSU HCSD) [11]. Public health surveillance data were used to identify PLWH out of care (i.e., no HIV viral load or CD4 count reported in the past 12 months). When patients accessed care at any LSU HCSD location, their name and demographics were matched with the out-of-care list. For out-of-care patients, a real-time EMR-based alert with clinical decision support was sent to the provider to prompt them to re-engage the patient in care.

More recently, public health departments have placed greater emphasis on data sharing as a strategy to improve relinkage to care through Data to Care initiatives [12, 13]. Through Data to Care, HIV care providers share their list of out-of-care patients with public health departments. Public health departments then match this "out-of-care" list with HIV surveillance data and send data back to HIV care providers regarding whether these patients are in care elsewhere. By forming this feedback loop, both public health departments and HIV care providers can improve the quality of their HIV surveillance and care data and target relinkage resources toward patients who are truly out of care, rather than those who have moved or transferred care [14].

Data to Care initiatives identify patients in need of relinkage after they have fallen out of care, but recent studies have focused on developing predictive models to identify PLWH at risk for retention in care failure *before* they disengage from care [15, 16]. Ramachandran et al. utilized EMR data combined with geospatial features and American Community Survey data to create a machine learning system to predict retention in care in an urban HIV clinic [15]. They compared the performance of various machine learning models including random forest models and logistic regression. Random forest is a machine learning method that combines the output from decision trees that are individually trained using sub-samples of data and features. The final prediction is made using the average of all tree predictions for regression models or using a majority vote for classification models. Ramachandran's study found that a random forest model had higher positive predictive value for flagging the top 10% highest risk patients compared to a logistic regression model [15]. Predicting retention in care in PLWH can also be done using unstructured text. Oliwa et al. used NLP of clinical notes to create a retention in care prediction model among PLWH [17]. They found that certain phrases within texts of notes, such as "syphilis," "K103N," "substance abuse," and

"stigma" were predictive of future lack of retention in care. Such models could be implemented within an HIV care clinic to allow retention resources to be directed toward patients most at risk for retention failure.

Viral Suppression

PLWH who achieve viral suppression with antiretroviral therapy experience improved health outcomes and are no longer able to transmit HIV to others. Thus, several studies have used EMR data to identify risk factors for virologic failure and to develop viral suppression prediction models [18–21]. Robbins et al. developed and validated a 1-year virologic failure prediction model using EMR data [21]. They then converted their model into a clinical prediction rule that providers can utilize to understand risk factors for virologic failure. The clinical prediction rule includes variables such as prior viral load, CD4 count, ART regimen, drug and alcohol abuse, and missed visits [21]. A recent study by Semerdjian et al. utilized NLP of clinical notes to predict HIV outcomes including viral suppression [22]. They found that a model using NLP of clinical notes had higher performance than a model based on demographics (AUC 0.83 vs. 0.75). Words/phrases found to be predictive of viral suppression included "migraine," "verruca," and "negative anxiety." Some of these NLP-detected terms may not seem to have a clinical association with viral suppression; i.e., it is unclear why a patient with migraines would be more or less likely to be virally suppressed than a patient without migraines. However, it is important to note that NLP algorithms do not necessarily detect that a patient has a certain condition, but only documentation of the condition in the clinical notes. It may be that providers who perform a detailed medical history and take the time to discuss minor medical conditions such as migraines with their patients are more likely to provide ART adherence support and resources for patients to facilitate viral suppression.

In addition to identifying risk factors for virologic failure or predicting viral suppression, other studies have investigated how clinical informatics interventions can be implemented to improve rates of viral suppression. Puttkamer et al. developed a prediction model for viral suppression including predictors such as consistency of ART medication pickups as well as clinical and social factors [23]. They then calculated a risk score and classified patients based on risk for future treatment failure. They incorporated the risk score into a best practice alert within the EMR to inform providers of patients' medication adherence and treatment failure risk. Providers received training for counseling at risk patients about medication adherence. The EMR alert and associated counseling were associated with a 15% greater likelihood of achieving viral suppression for patients who received the intervention.

Several additional studies have used machine learning models to not only predict virologic failure but to also determine the optimal intervals at which viral load tests should be collected [24–26]. Petersen et al. used the super learner machine learning algorithm with medication event monitoring systems (MEMS) data to develop a model of virologic failure [26]. The model was then used to predict the proportion of HIV viral load tests that could have been avoided based on the probability that they would have shown viral suppression. The study found that 25–31% of viral loads could have been avoided, allowing for savings of $16–$29 dollars per person-month.

Challenges in Utilizing EMR Data and Machine Learning for HIV Care

Inconsistent EMR Documentation

The development of reliable EMR algorithms depends on the presence of accurate information within the EMR. Unfortunately, EMRs often contain incorrect or missing documentation of factors relevant for HIV care. For example, despite recommendations from the National Academy of Medicine and the Joint Commission [27–29], many EMRs do not have a systematic way for documenting sexual orientation or gender identity. PLWH are disproportionately impacted by psychiatric illness compared to the general population. However, Brown et al. found that psychiatric illness and substance use disorder are under-documented in structured fields in EMR records for PLWH [30].

One strategy to overcome EMR under-documentation is to utilize algorithms that incorporate multiple EMR fields for relevant conditions. For example, to identify patients with psychiatric illness, an algorithm could take into account diagnostic codes, documentation of mental illness in the problem list or past medical history, mental health screening results, prescription of psychiatric medications, or clinical encounters in the Psychiatry department. Moreover, use of natural language processing can detect factors present in clinical notes that are not documented in structured EMR fields. Ridgway et al. found that among patients with psychiatric illness or substance use disorder detected by NLP of clinical notes, only half had these behavioral health disorders documented in structured EMR fields [31].

Even something as foundational as identifying people who have tested positive for HIV may require multi-step algorithms due to incomplete EMR data. Paul et al. developed two EMR-based algorithms that included HIV antibody test results, viral load test results, antiretroviral therapy prescriptions, and ICD-9 codes [32]. Their algorithms had high specificity of 99–100% but lower sensitivity of 77–78% for accurately identifying PLWH within an EMR database. The most

common reasons for the algorithms failing to identify PLWH were missing laboratory or medication data from the EMR and patients being diagnosed with HIV at an outside institution.

Barriers to Data Sharing

An additional challenge related to use of EMR data for HIV patient care and epidemiologic and clinical research is the lack of data sharing between healthcare systems. PLWH often receive care at different healthcare facilities and may have laboratory results and/or clinical notes in different healthcare systems' EMRs that may not be linked. Healthcare data fragmented in disparate EMR systems results in a lack of a complete clinical picture at any given healthcare site.

Data sharing among healthcare organizations requires significant resources such as informatics support for harmonizing data across different platforms. There are also data security considerations and protections that must be in place to support privacy and confidentiality of data, particularly related to HIV status which is highly sensitive health information. Permissions for data sharing may not be uniform across healthcare systems, and public health institutions may have policies against disclosing HIV data to clinical entities. Moreover, healthcare systems frequently update or change their EMR systems, and processes for data sharing must be continually maintained through these updates. Despite these challenges, several groups have formed data sharing platforms for HIV data, such as the LaPHIE HIV care system to improve retention in care that was previously mentioned [11]. Several EMR-based HIV research cohorts have also been developed with data from multiple HIV care sites. These include the Center for AIDS Research Center Network of Integrated Clinical Systems (CNICS) cohort and the DC cohort [33, 34]. Such cohorts require resources and commitment from all participating sites as well as continued funding to support ongoing collaboration. More support and incentives are needed to facilitate data sharing for PLWH to improve quality of data and ultimately quality of care for PLWH.

Challenges in Utilizing Clinical Decision Support Tools

Clinical decision support tools have the potential to improve care for PLWH by guiding providers regarding care for their patients. However, it can be challenging to build and implement these tools. Healthcare systems may have competing priorities and may prioritize other EMR tools over those for HIV care. Moreover, providers may not utilize the tools or respond to the alerts. "Alert fatigue" describes the phenomenon in which providers become desensitized to repeated alerts in the EMR, prompting them to override and ignore such alerts [35, 36]. In clinical practice, the majority of CDS alerts are overridden, thereby limiting their utility [37].

To successfully improve care, CDS tools must follow clinical informatics best practices (i.e., fitting into the provider's workflow and minimizing extra "clicks") [38]. Alerts that do not follow these best practices will likely fail to improve HIV care. For example, one institution found their EMR HIV screening alert to be ineffective because it prompted providers to enter documentation of patients' verbal consent for HIV testing *after* providers had finished speaking with patients. Because this alert failed to fit into providers' workflow, it was ignored over 99% of the time [39]. Similarly, important metrics for prediction such as number needed to screen, sensitivity, and positive predictive values must be reported by studies that build prediction models.

Potential for Bias in Machine Learning

Machine learning algorithms have enormous potential for improving HIV care but can also pose additional challenges. Although algorithms may avoid biases in diagnosis and treatment by objectively synthesizing and analyzing data, they can also perpetuate bias among historically marginalized communities, many of whom are disproportionately affected by HIV [40]. Machine learning algorithms used for risk prediction can reflect human biases in decision-making and exhibit substantial racial or gender bias, inadvertently perpetuating or exacerbating health disparities [40–44]. Bias within machine learning in healthcare can exist in the design, data, and deployment of a model [43] and is usually associated with missing data and certain groups or individual patients not being identified by algorithms, sample size underestimation, and misclassification and measurement error [40]. For instance, machine learning models utilize historically collected data, meaning that vulnerable groups who have endured human and structural biases are subject to harm by either incorrect predictions or withholding of certain resources [43].

HIV disproportionately impacts Blacks/African Americans, who account for a higher proportion of new HIV diagnoses compared to people of other races/ethnicities and are most vulnerable to machine learning bias [45]. A study assessing algorithm performance for HIV risk prediction found that a majority of the machine learning models based on variables related to sexual orientation and STIs had lower sensitivity for Black patients than White patients [46]. This disparity could result from a lack of traditional HIV risk factors documented within the medical records of Black people due to factors such as stigma and medical mistrust [47] and structural racism within the healthcare system [48] that can impact the accuracy of the information within their medical records [49]. In addition to race, the study evaluated predictive performance by sex and found none of the algorithms used in their healthcare setting predicted HIV acquisition among women, further demonstrating bias inherent within their machine learning algorithms [46].

When machine learning algorithms are biased, they can further perpetuate inequalities. Machine learning algorithms for HIV care must be developed, implemented, and evaluated with principles of distributive justice [43]. The investigators who design machine learning algorithms must understand and address potential biases, such as structural racism, misogyny, and discrimination against sexual and gender minorities, and ensure algorithms will advance health equity and benefit all patients [50]. Strategies to address and overcome bias in machine learning include engaging various stakeholders in the design and implementation process, measuring algorithm performance across diverse groups, and monitoring patient outcomes [15, 43]. Properly designed and utilized machine learning could help to resolve disparities in healthcare, especially those related to the HIV epidemic, if algorithms remedy known biases and highlight areas for future research [42].

Future Directions and Innovations

While EMR data are a rich source of information regarding PLWH, they are limited in their ability to identify social and structural factors that impact HIV care. The vast majority of people's lives are spent outside the healthcare system, and EMR data only offer a snapshot of factors that impact overall health. People generate enormous quantities of data outside of the EMR in their daily lives through social media, internet searches, geospatial tracking, etc. Future work should seek to supplement EMR data with these additional data sources.

Research to date on the use of these additional data sources for HIV-related data has been promising. One study used machine learned methods to examine patterns in HIV risk behavior documented on Twitter and found that their models were able to identify HIV-related Tweets with a mean accuracy of 85% [51]. Young et al. found that Facebook data including Facebook group affiliations and social network structures are associated with sex behaviors that may impact HIV transmission [52, 53]. Others have used internet search query data to predict locations of new HIV diagnoses in China [54, 55]. Use of social media data for HIV-related research poses unique ethical considerations given that social media companies may use or sell individuals' personal data for profit. When social media data are combined with health data, extra precautions must be taken to ensure individuals' privacy.

There has also been a recent recognition that geographic and neighborhood factors may influence HIV care continuum outcomes. At the individual level, geospatial analyses have shown that longer travel time to HIV clinic is associated with decreased retention in care [56]. Global Positioning System (GPS) technology can be utilized to understand mobility within neighborhoods and access to needed resources among people living with or vulnerable to HIV [57]. At the community level, geospatial analyses have shown that there exist geographic "hot spots" wherein PLWH are less likely to be retained in care or virally suppressed [58]. Community characteristics such as lower walkability scores and more vacant buildings have been associated with increased incidence of HIV infection [59]. Individuals' addresses can be mapped onto community level data, including neighborhood and socioeconomic data from the American Community Survey, crime rates, rates of sexually transmitted infections, and other public health data to better understand factors associated with HIV care continuum outcomes among PLWH.

While there has been limited research to date combining these various data sources with EMR data for HIV care continuum research, several promising studies are underway. In South Carolina, one group is creating a database of PLWH linking surveillance data from the state health department with EMR data, crime and prison data from the Department of Corrections, mental health data, and socioeconomic data from American Community Survey [16]. They plan to use machine learning techniques to characterize and predict HIV care continuum outcomes using the database.

Recent research has moved from using EMR data and other sources for *descriptive analytics*, i.e., describing and understanding patterns in HIV epidemiology, toward *predictive analytics*, i.e., predicting which patients are most likely to experience poor HIV care continuum outcomes. More research is needed to understand the best ways to utilize these predictive models in practice. It is not known how these models can best fit into care teams' workflows and how they can complement current tools and practices as well as providers' own intuition regarding their patients' likely outcomes. Implementation science methods should be utilized to guide their use. Research is also needed to understand the perspectives of PLWH regarding the use of their personal health data for predictive modeling, including concerns about privacy and bias.

Conclusions

With the increase in utilization of EMRs and the application of machine learning methods, EMR data are a rich data source for expanding HIV-related knowledge. Predictive analytic techniques combined with clinical informatics offer the potential for medical providers to intervene in real time to improve HIV care continuum outcomes for at risk patients, from diagnosis to viral suppression. Working with EMR data does have challenges, including missing documentation, difficulty harmonizing data from different EMR systems, and privacy and confidentiality concerns. Moreover, machine learning methods can exacerbate disparities by perpetuating bias, and researchers must analyze and correct for potential bias in their models. Despite these challenges, research to date has highlighted the promise of these technologies. Promising

future areas of research include incorporating HIV-related EMR data with other social and structural data sources, such as social media data, geospatial data, and public health data. More research is also needed to understand the best way to implement HIV-related predictive models into clinical care for PLWH to improve care across the HIV care continuum.

Declarations

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

Human and Animal Rights and Informed Consent   This article does not contain any studies with human or animal subjects performed by any of the authors.

References

1. The Office of the National Coordinator for Health Information Technology. Health IT Dashboard. https://dashboard.healthit.gov/apps/health-information-technology-data-summaries.php?state=National&cat9=all+data&cat1=ehr+adoption#summary-data. Accessed November 29, 2020. .

2. Ahlstrom MG, Ronit A, Omland LH, Vedel S, Obel N. Algorithmic prediction of HIV status using nation-wide electronic registry data. EClinicalMedicine. 2019;17:100203. https://doi.org/10.1016/j.eclinm.2019.10.016.

3. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. J Acquir Immune Defic Syndr. 2018;77(2):160–6. https://doi.org/10.1097/QAI.0000000000001580.

4. Force UPST. Screening for HIV infection: US preventive services task force recommendation statement. JAMA. 2019;321(23):2326–36. https://doi.org/10.1001/jama.2019.6587.

5. Branson BM, Handsfield HH, Lampe MA, Janssen RS, Taylor AW, Lyss SB, et al. Revised recommendations for HIV testing of adults, adolescents, and pregnant women in health-care settings. MMWR Recomm Rep. 2006;55(RR-14):1–17 quiz CE1-4.

6. Dailey AF, Hoots BE, Hall HI, Song R, Hayes D, Fulton P Jr, et al. Vital signs: human immunodeficiency virus testing and diagnosis delays - United States. MMWR Morb Mortal Wkly Rep. 2017;66(47):1300–6. https://doi.org/10.15585/mmwr.mm6647e1.

7. Lin J, Mauntel-Medici C, Heinert S, Baghikar S. Harnessing the power of the electronic medical record to facilitate an opt-out HIV screening program in an urban academic emergency department. J Public Health Manag Pract. 2017;23(3):264–8. https://doi.org/10.1097/PHH.0000000000000448.

8. Marcelin JR, Tan EM, Marcelin A, Scheitel M, Ramu P, Hankey R, et al. Assessment and improvement of HIV screening rates in a Midwest primary care practice using an electronic clinical decision support system: a quality improvement study. BMC Med Inform Decis Mak. 2016;16:76. https://doi.org/10.1186/s12911-016-0320-5.

9. Ridgway JP, Almirol E, Schmitt J, Wesley-Madgett L, Pitrak D. A Clinical informatics approach to reengagement in HIV care in the emergency department. J Public Health Manag Pract. 2019;25(3):270–3. https://doi.org/10.1097/PHH.0000000000000844.

10. Shade SB, Steward WT, Koester KA, Chakravarty D, Myers JJ. Health information technology interventions enhance care completion, engagement in HIV care and treatment, and viral suppression among HIV-infected patients in publicly funded settings. J Am Med Inform Assoc. 2015;22(e1):e104–11. https://doi.org/10.1136/amiajnl-2013-002623.

11. Herwehe J, Wilbright W, Abrams A, Bergson S, Foxhood J, Kaiser M, et al. Implementation of an innovative, integrated electronic medical record (EMR) and public health information exchange for HIV/AIDS. J Am Med Inform Assoc. 2012;19(3):448–52. https://doi.org/10.1136/amiajnl-2011-000412.

12. Sweeney P, DiNenno EA, Flores SA, Dooley S, Shouse RL, Muckleroy S, et al. HIV Data to care-using public health data to improve HIV care and prevention. J Acquir Immune Defic Syndr. 2019;82(Suppl 1):S1–5. https://doi.org/10.1097/QAI.0000000000002059.

13. Centers for Disease Control and Prevention. Data to Care. . https://www.cdc.gov/hiv/effective-interventions/respond/data-to-care?Sort=Title%3A%3Aasc&Intervention%20Name=Data%20to%20Care. Accessed November 20, 2020.

14. Ridgway JP, Schmitt J, Almirol E, Millington M, Harding E, Pitrak D. Electronic data sharing between public health department and clinical providers improves accuracy of HIV retention data. Open Forum Infect Dis. 2017;4(Suppl 1):S421–2. Published 2017 Oct 4. https://doi.org/10.1093/ofid/ofx163.1059.

15. Ramachandran A, Kumar A, Koenig H, De Unanue A, Sung C, Walsh J, et al. Predictive analytics for retention in care in an urban HIV clinic. Sci Rep. 2020;10(1):6421. https://doi.org/10.1038/s41598-020-62729-x.

16. 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. https://doi.org/10.1136/bmjopen-2018-027688.

17. Oliwa T, Furner B, Schmitt J, Schneider J, Ridgway JP. Development of a predictive model for retention in HIV care using natural language processing of clinical notes. J Am Med Inform Assoc. 2020;28:104–12. https://doi.org/10.1093/jamia/ocaa220.

18. Dessie ZG, Zewotir T, Mwambi H, North D. Modeling viral suppression, viral rebound and state-specific duration of HIV patients with CD4 count adjustment: parametric multistate frailty model approach. Infect Dis Ther. 2020;9(2):367–88. https://doi.org/10.1007/s40121-020-00296-4.

19. Gebrezgi MT, Fennie KP, Sheehan DM, Ibrahimou B, Jones SG, Brock P, et al. Development and validation of a risk prediction tool to identify people with HIV infection likely not to achieve viral suppression. AIDS Patient Care STDs. 2020;34(4):157–65. https://doi.org/10.1089/apc.2019.0224.

20. Bisaso KR, Karungi SA, Kiragga A, Mukonzo JK, Castelnuovo B. A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating HIV patients. BMC Med Inform Decis Mak. 2018;18(1):77. https://doi.org/10.1186/s12911-018-0659-x.

21. Robbins GK, Johnson KL, Chang Y, Jackson KE, Sax PE, Meigs JB, et al. Predicting virologic failure in an HIV clinic. Clin Infect Dis. 2010;50(5):779–86. https://doi.org/10.1086/650537.

22. Semerdjian J, Lykopoulos K, Maas A, Harrell M, Priest J, Eitz-Ferrer P et al. Supervised machine learning to predict HIV outcomes using electronic health record and insurance claims data. AIDS 2018. 2018; http://programme.aids2018.org/Abstract/Abstract/4559.

23. Puttkammer N, Simoni JM, Sandifer T, Chery JM, Dervis W, Balan JG, et al. An EMR-based alert with brief provider-led ART adherence counseling: promising results of the InfoPlus adherence pilot study among Haitian adults with HIV initiating ART. AIDS Behav. 2020;24(12):3320–36. https://doi.org/10.1007/s10461-020-02945-8.

24. Kamal S, Urata J, Cavassini M, Liu H, Kouyos R, Bugnon O, et al. Random forest machine learning algorithm predicts virologic outcomes among HIV infected adults in Lausanne, Switzerland using electronically monitored combined antiretroviral treatment

adherence. AIDS Care. 2020:1–7. https://doi.org/10.1080/09540121.2020.1751045.

25. Benitez AE, Musinguzi N, Bangsberg DR, Bwana MB, Muzoora C, Hunt PW, et al. Super learner analysis of real-time electronically monitored adherence to antiretroviral therapy under constrained optimization and comparison to non-differentiated care approaches for persons living with HIV in rural Uganda. J Int AIDS Soc. 2020;23(3):e25467. https://doi.org/10.1002/jia2.25467.

26. Petersen ML, LeDell E, Schwab J, Sarovar V, Gross R, Reynolds N, et al. Super learner analysis of electronic adherence data improves viral prediction and may provide strategies for selective HIV RNA monitoring. J Acquir Immune Defic Syndr. 2015;69(1):109–18. https://doi.org/10.1097/QAI.0000000000000548.

27. National Academy of Medicine. The health of lesbian, gay, bisexual, and transgender (LGBT) people: building a foundation for better understanding. Washington, DC: National Academies Press; 2011. Available from: www.nap.edu/catalog.php?record_id=13128external.

28. Deutsch MB, Green J, Keatley J, Mayer G, Hastings J, Hall AM, et al. Electronic medical records and the transgender patient: recommendations from the World Professional Association for Transgender Health EMR Working Group. J Am Med Inform Assoc. 2013;20(4):700–3. https://doi.org/10.1136/amiajnl-2012-001472.

29. Deutsch MB, Buchholz D. Electronic health records and transgender patients–practical recommendations for the collection of gender identity data. J Gen Intern Med. 2015;30(6):843–7. https://doi.org/10.1007/s11606-014-3148-7.

30. Brown LA, Mu W, McCann J, Durborow S, Blank MB. Under-documentation of psychiatric diagnoses among persons living with HIV in electronic medical records. AIDS Care. 2020;33:1–5. https://doi.org/10.1080/09540121.2020.1713974.

31. Ridgway J, Uvin A, Schmitt J, Oliwa T, Almirol E, Devlin S, et al. Natural language processing of clinical notes to identify mental illness and substance use among people living with HIV. JMIR Med Inform (forthcoming. https://doi.org/10.2196/23456.

32. Paul DW, Neely NB, Clement M, Riley I, Al-Hegelan M, Phelan M, et al. Development and validation of an electronic medical record (EMR)-based computed phenotype of HIV-1 infection. J Am Med Inform Assoc. 2018;25(2):150–7. https://doi.org/10.1093/jamia/ocx061.

33. Greenberg AE, Hays H, Castel AD, Subramanian T, Happ LP, Jaurretche M, et al. Development of a large urban longitudinal HIV clinical cohort using a web-based platform to merge electronically and manually abstracted data from disparate medical record systems: technical challenges and innovative solutions. J Am Med Inform Assoc. 2016;23(3):635–43. https://doi.org/10.1093/jamia/ocv176.

34. Kitahata MM, Rodriguez B, Haubrich R, Boswell S, Mathews WC, Lederman MM, et al. Cohort profile: the centers for AIDS research network of integrated clinical systems. Int J Epidemiol. 2008;37(5):948–55. https://doi.org/10.1093/ije/dym231.

35. Ancker JS, Edwards A, Nosal S, Hauser D, Mauer E, Kaushal R, et al. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC Med Inform Decis Mak. 2017;17(1):36. https://doi.org/10.1186/s12911-017-0430-8.

36. Embi PJ, Leonard AC. Evaluating alert fatigue over time to EHR-based clinical trial alerts: findings from a randomized controlled study. J Am Med Inform Assoc. 2012;19(e1):e145–8. https://doi.org/10.1136/amiajnl-2011-000743.

37. Isaac T, Weissman JS, Davis RB, Massagli M, Cyrulik A, Sands DZ, et al. Overrides of medication alerts in ambulatory care. Arch Intern Med. 2009;169(3):305–11. https://doi.org/10.1001/archinternmed.2008.551.

38. Bates DW, Kuperman GJ, Wang S, Gandhi T, Kittler A, Volk L, et al. Ten commandments for effective clinical decision support: making the practice of evidence-based medicine a reality. J Am Med Inform Assoc. 2003;10(6):523–30. https://doi.org/10.1197/jamia.M1370.

39. Kao C. Personal communication. March. 2017;22.

40. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. JAMA Intern Med. 2018;178(11):1544–7. https://doi.org/10.1001/jamainternmed.2018.3763.

41. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. https://doi.org/10.1126/science.aax2342.

42. Char DS, Shah NH, Magnus D. Implementing machine learning in health care - addressing ethical challenges. N Engl J Med. 2018;378(11):981–3. https://doi.org/10.1056/NEJMp1714229.

43. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018;169(12):866–72. https://doi.org/10.7326/M18-1990.

44. Marcus JL, Sewell WC, Balzer LB, Krakower DS. Artificial intelligence and machine learning for HIV prevention: emerging approaches to ending the epidemic. Curr HIV/AIDS Rep. 2020;17(3):171–9. https://doi.org/10.1007/s11904-020-00490-6.

45. Prevention CfDCa. HIV and African Americans. 2020. https://www.cdc.gov/hiv/group/racialethnic/africanamericans/index.html. Accessed November 6 2020.

46. Marcus JL, Hurley LB, Hare CB, Silverberg MJ, Volk JE. Disparities in uptake of HIV preexposure prophylaxis in a large integrated health care system. Am J Public Health. 2016;106(10):e2–3. https://doi.org/10.2105/AJPH.2016.303339.

47. Eaton LA, Driffin DD, Kegler C, Smith H, Conway-Washington C, White D, et al. The role of stigma and medical mistrust in the routine health care engagement of black men who have sex with men. Am J Public Health. 2015;105(2):e75–82. https://doi.org/10.2105/AJPH.2014.302322.

48. Feagin J, Bennefield Z. Systemic racism and U.S. health care. Soc Sci Med. 2014;103:7–14. https://doi.org/10.1016/j.socscimed.2013.09.006.

49. Klinger EV, Carlini SV, Gonzalez I, Hubert SS, Linder JA, Rigotti NA, et al. Accuracy of race, ethnicity, and language preference in an electronic health record. J Gen Intern Med. 2015;30(6):719–23. https://doi.org/10.1007/s11606-014-3102-8.

50. Robinson WR, Renson A, Naimi AI. Teaching yourself about structural racism will improve your machine learning. Biostatistics. 2020;21(2):339–44. https://doi.org/10.1093/biostatistics/kxz040.

51. 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. 2017;74(Suppl 2):S128–S31. https://doi.org/10.1097/QAI.0000000000001240.

52. Young LE, Fujimoto K, Schneider JA. HIV Prevention and sex behaviors as organizing mechanisms in a facebook group affiliation network among young black men who have sex with men. AIDS Behav. 2018;22(10):3324–34. https://doi.org/10.1007/s10461-018-2087-4.

53. Young LE, Ramachandran A, Schumm LP, Khanna AS, Schneider JA. The potential of online social networking data for augmenting the study of high-risk personal networks among young men who have sex with men at-risk for HIV. Soc Networks. 2020;63:201–12. https://doi.org/10.1016/j.socnet.2020.06.003.

54. Zhang Q, Chai Y, Li X, Young SD, Zhou J. Using internet search data to predict new HIV diagnoses in China: a modelling study. BMJ Open. 2018;8(10):e018335. https://doi.org/10.1136/bmjopen-2017-018335.

55. Nan Y, Gao Y. A machine learning method to monitor China's AIDS epidemics with data from Baidu trends. PLoS One. 2018;13(7):e0199697. https://doi.org/10.1371/journal.pone.0199697.

56. Ridgway JP, Almirol EA, Schmitt J, Schuble T, Schneider JA. Travel time to clinic but not neighborhood crime rate is associated with retention in care among HIV-positive patients. AIDS Behav. 2018;22(9):3003–8. https://doi.org/10.1007/s10461-018-2094-5.

57. Duncan DT, Hickson DA, Goedel WC, Callander D, Brooks B, Chen YT, et al. The social context of HIV prevention and care among black men who have sex with men in three U.S. cities: the neighborhoods and networks (N2) cohort study. Int J Environ Res Public Health. 2019;16(11). https://doi.org/10.3390/ijerph16111922.

58. Eberhart MG, Yehia BR, Hillier A, Voytek CD, Blank MB, Frank I, et al. Behind the cascade: analyzing spatial patterns along the HIV care continuum. J Acquir Immune Defic Syndr. 2013;64(Suppl 1): S42–51. https://doi.org/10.1097/QAI.0b013e3182a90112.

59. Phillips G 2nd, Birkett M, Kuhns L, Hatchel T, Garofalo R, Mustanski B. Neighborhood-level associations with HIV infection among young men who have sex with men in Chicago. Arch Sex Behav. 2015;44(7):1773–86. https://doi.org/10.1007/s10508-014-0459-z.