Published in final edited form as:
Lancet HIV. 2019 October ; 6(10): e688–e695. doi:10.1016/S2352-3018(19)30137-7.

Use of electronic health record data and machine learning to identify potential candidates for HIV preexposure prophylaxis: a modelling study

Julia L. Marcus, PhD,
Harvard Medical School and Harvard Pilgrim Health Care Institute, 401 Park Drive, Boston, Massachusetts, 02215, USA

Leo B. Hurley, MPH,
Kaiser Permanente Division of Research, 2000 Broadway, Oakland, California, 94612, USA

Douglas S. Krakower, MD,
Harvard Medical School and Harvard Pilgrim Health Care Institute, 401 Park Drive, Boston, Massachusetts, 02215, USA; Beth Israel Deaconess Medical Center, 330 Brookline Avenue, Boston, Massachusetts, 02215, USA

Stacey Alexeeff, PhD,
Kaiser Permanente Division of Research, 2000 Broadway, Oakland, California, 94612, USA

Michael J. Silverberg, PhD,
Kaiser Permanente Division of Research, 2000 Broadway, Oakland, California, 94612, USA

Jonathan E. Volk, MD
Kaiser Permanente San Francisco Medical Center, 2238 Geary Boulevard, San Francisco, California, 94115, USA

SUMMARY

Background.—The limitations of existing HIV risk prediction tools are a barrier to preexposure prophylaxis (PrEP) implementation. We developed and validated an HIV prediction model to identify potential PrEP candidates in a large healthcare system.

Corresponding author: Julia L. Marcus, PhD, MPH, 401 Park Dr, Boston, MA 02215, Tel: (617) 867-4557, julia_marcus@harvardpilgrim.org.

AUTHORS' CONTRIBUTIONS

JLM and JEV conceived the study and obtained funding. All authors made substantial contributions to the design of the study. LBH collected the data. JLM analyzed the data and drafted the manuscript. SA provided biostatistical support. All authors contributed to interpretation of data, revised the manuscript for important intellectual content, and approved the final version for publication.

Previous presentation: This work was presented as an oral abstract at the Conference on Retroviruses and Opportunistic Infections, March 4-7, 2019, Seattle, Washington.

DECLARATION OF INTERESTS

JLM has received research grant support from Merck and has consulted on a research grant to Kaiser Permanente Northern California from Gilead Sciences, both outside the submitted work. DSK has conducted research with project support from Gilead Sciences, has received honoraria for authoring or presenting continuing medical education content for Medscape, MED-IQ, and DKBMed, and has received royalties for authoring content for Uptodate, Inc. MJS has received research grant support from Merck and Gilead Sciences, both outside the submitted work. SA has received research grant support from Gilead Sciences outside the submitted work. All other authors report no conflicts.

Methods.—Our study population was HIV-uninfected members of Kaiser Permanente Northern California not yet using PrEP. Using 81 electronic health record (EHR) variables, we applied least absolute shrinkage and selection operator (LASSO) regression to predict incident HIV diagnosis within three years in 2007-2014, assessing ten-fold cross-validated area under the curve (C-statistic). We compared the full model to simpler models including only men who have sex with men (MSM) status and sexually transmitted infection (STI) positivity. Models were validated prospectively with 2015-2017 data.

Findings.—Of 3,750,664 patients in 2007-2017, there were 784 incident HIV cases within three years. The LASSO procedure retained 44 predictors in the full model, with a C-statistic of 0·86 (95% CI: 0·85-0·87) for incident HIV cases in 2007-2014. Model performance remained high in 2015-2017 (C-statistic 0·84, 95% CI: 0·80-0·89). The full model outperformed simpler models including only MSM status and STI positivity. For the full model, flagging patients with high or very high HIV risk scores in the validation dataset (13,463/606,701, 2·2%) identified 38·6% (32/83) of the incident HIV cases, including 46·4% (32/69) of male cases and 0% (0/14) of female cases. The full model had equivalent sensitivity by race, while simpler models identified fewer Black than White HIV cases.

Interpretation.—Prediction models using EHR data can identify patients at high risk of HIV acquisition but not yet using PrEP. Future studies should optimize EHR-based HIV risk prediction tools and evaluate their impact on PrEP prescribing.

Funding.—Kaiser Permanente Community Benefit Research Program and the US National Institutes of Health.

BACKGROUND

There are nearly 40,000 new HIV infections annually in the U.S. Preexposure prophylaxis (PrEP) using co-formulated tenofovir disoproxil fumarate and emtricitabine is over 90% effective in preventing HIV acquisition and is recommended by the U.S. Centers for Disease Control and Prevention (CDC) for populations at high risk of acquiring HIV infection.[1] However, the CDC estimates that only seven percent of the 1·1 million individuals in the U.S. with indications for PrEP used it in 2016.[2]

One barrier to PrEP implementation, as noted by the U.S. Preventive Services Task Force, is the challenge of identifying individuals who may benefit from PrEP.[3] Although the CDC has specified indications for PrEP, including having multiple sex partners or a recent bacterial sexually transmitted infection (STI),[1] healthcare providers report difficulty identifying patients at risk of HIV acquisition.[4] Providers have time constraints, may be uneasy conducting sexual and substance use histories, and may believe few of their patients have indications for PrEP.[4,5] Existing HIV risk prediction tools require that providers already know a patient is in a particular risk group, such as men who have sex with men (MSM).[3,6,7] Moreover, existing prediction tools based on CDC criteria for PrEP use have been shown to underestimate risk for HIV acquisition in Black MSM.[8,9] Risk prediction tools that automate identification of patients at high risk of HIV acquisition in general patient populations could support providers in more efficient and equitable assessment of their patients' suitability for PrEP.

Our objective was to develop and validate an HIV risk prediction model using electronic health record (EHR) data collected during routine clinical care to identify potential PrEP candidates in a large healthcare system in California. To understand the extent to which multiple EHR data domains can improve identification of PrEP candidates beyond traditional HIV risk factors, including those noted as indications for PrEP in CDC guidelines,[1] we compared the performance of a full model to models based on only MSM status and recent bacterial STIs. Finally, we assessed the ability of models to identify potential PrEP candidates in subpopulations, specifically females and Black individuals, that have not been fully captured by existing HIV risk prediction tools.

METHODS

Study Setting, Population, and Design

Our study setting was Kaiser Permanente Northern California (KPNC), a large integrated healthcare system that provides comprehensive medical services to 4·3 million members, corresponding to approximately 30% of insured individuals in the surrounding population.[10] We developed and validated models to predict incident HIV diagnosis within three years among all adult (aged ≥18) KPNC patients who had at least two years of prior health plan enrollment with at least one outpatient visit during 2007-2017. We excluded patients who had been diagnosed with HIV infection prior to baseline, and also excluded those who had a prior pharmacy fill for PrEP. We determined dates of first lifetime HIV diagnosis using the KPNC HIV registry, which includes all known HIV cases since the early 1980s, with cases confirmed by chart review. The start of follow-up for each subject (baseline) was the earliest date on or after January 1, 2007, when eligibility criteria were met. Subjects were followed until the earliest of HIV diagnosis, disenrollment from the health plan, death, or December 31, 2017.

The institutional review board at KPNC approved this study with a waiver of written informed consent.

Predictors

We extracted demographic and clinical data from KPNC's EHR, yielding 81 potential predictors of HIV risk that were identified based on published literature and clinical expertise (Appendix pages 2-3). Most variables had no missing data. For example, the presence of a given diagnosis was coded as 1 and the absence of that diagnosis as 0. To define MSM status, we coded anyone who was male and ever previously reported male sex partners as 1; everyone else was coded as 0, even those who had missing data on sexual orientation.

Model Development

Model development and validation followed the guidelines for Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD).[11] We developed models using the subset of patients who entered the cohort in 2007-2014 (development dataset). Using EHR variables from prior to each subject's baseline date, we fitted models to predict incident HIV diagnosis within three years. We used least absolute shrinkage and selection operator (LASSO) logistic regression, which automates selection of a subset of variables to optimize predictive accuracy,[12,13] with weighting to account for differences in duration of follow-up.[14]

Model discrimination was assessed by using the C-statistic, which represents the probability that a randomly drawn HIV case was ranked as higher risk by the model than a randomly drawn non-case; the C-statistic can be computed as the area under the receiver operating characteristic curve.[15] Given that prediction models might perform well in the development dataset but not generalize to new data if they are overfitted, we used ten-fold cross-validation to minimize overfitting.[12] We randomly split the development dataset into ten equally sized subsamples, with nine used to develop the model and the remaining subsample used for testing the model; this partitioning was repeated ten times (folds). We averaged the results across folds to estimate the cross-validated C-statistic.

We developed both a full model, which included all potential predictors, and simpler models that included only MSM status and/or STI variables. We evaluated STI positivity, a set of five variables indicating numbers of positive tests for gonorrhea and chlamydia by anatomic site, and reactive syphilis tests, in the prior two years. We also evaluated four additional variables indicating a history of rectal STI testing, syphilis testing, and syphilis treatment.

Model Validation

To assess how the full and simpler models might perform prospectively, we validated them using data from an independent set of patients who entered the cohort in 2015-2017 (validation dataset), which was not used during any part of the model development process. We again computed C-statistics to assess model discrimination, generating 95% confidence intervals (CIs) using bootstrapping with 1000 resamples of the data.

Identification of PrEP Candidates

We computed predicted probabilities of incident HIV diagnosis within three years, or HIV risk scores, for all patients in the 2015-2017 validation dataset given their values of the predictors retained in the models. We categorized risk scores as follows: low (<·05%), moderate (0·05% - <0·20%), high (0·20% - <1·0%), and very high (≥1·0%). To our knowledge, there are no established thresholds of predicted HIV risk for determining PrEP eligibility. For context, we compared the observed HIV risk in subgroups likely to have indications for PrEP, specifically patients with bacterial STIs in the prior two years and MSM, with the observed risk in the general patient population. These subgroups were at approximately 10 and 50 times higher risk than the average patient, respectively, corresponding with the categories of high and very high risk scores. For the full model, calibration (i.e., agreement between observed outcomes and predictions) was assessed by plotting observed and predicted probabilities of incident HIV diagnosis within three years across categories of predicted risk.

We then assessed the sensitivity of each model for identification of incident HIV cases, defined as the proportion of incident HIV cases identified among patients with high or very high HIV risk scores, overall and by sex and race. Finally, to understand whether model predictions matched real-world clinical judgment, we assessed sensitivity for identification of patients who were actually prescribed PrEP, defined as the proportion of PrEP users identified among patients with high or very high risk scores.

Analyses were conducted in the R environment for statistical computing, version 3.5.1.

Role of the Funding Source

The funder of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

RESULTS

Of 3,761,028 adult KPNC patients who had at least two years of prior health plan enrollment and an outpatient visit from January 1, 2007, through December 31, 2017, a total of 9290 were excluded because of a prior HIV diagnosis and 1074 because of a prior pharmacy fill for PrEP at KPNC. Of the 3,750,664 HIV-uninfected KPNC patients not yet using PrEP, there were 784 incident HIV diagnoses occurring within three years of baseline. Characteristics of patients in the development and validation datasets are shown in Table 1.

The LASSO variable selection procedure retained 44 predictors in the full model, with a cross-validated C-statistic of 0·86 (95% CI: 0·85-0·87) for identification of incident HIV cases in the 2007-2014 dataset. The HIV risk predictors retained in the full model, their prevalence among patients with and without an incident HIV diagnosis, and adjusted odds ratios in the 2007-2014 dataset are shown in Table 2. The strongest predictors were male sex, MSM status, residing in a ZIP code with high HIV incidence, urine positivity for methadone, number of positive tests for urethral gonorrhea in the prior two years, and number of penicillin G benzathine injections with a syphilis test within 90 days in the prior two years. The prevalence and adjusted odds ratios for HIV risk predictors retained in simpler models are shown in the Appendix (page 4).

Model discrimination remained high when validated prospectively, with a C-statistic of 0·84 (95% CI: 0·80-0·89) in the 2015-2017 dataset (Figure 1). Discrimination was substantially reduced when including only MSM status and STI positivity, MSM status alone, or STI positivity alone. Incorporating additional variables for STI testing and treatment improved discrimination of these simpler models.

Based on the full model, a small subset of patients in the 2015-2017 dataset had high (11,930/606,701, 2·0%) or very high HIV risk scores (1533/606,701, 0·3%). The full model was well-calibrated for patients with low, moderate, and high risk scores, but overestimated the absolute risk of incident HIV diagnosis within three years among patients with very high risk scores (data not shown).

We assessed sensitivity for each model (Table 3). Flagging patients with high or very high HIV risk scores from the full model identified 38·6% (32/83) of the incident HIV cases overall, including 46·4% (32/69) of the cases among males. Flagging only the patients with very high risk scores identified 15·7% (13/83) of incident HIV cases. Simpler models flagged fewer patients and had lower sensitivity for incident HIV cases. No models identified incident HIV cases among females. The full model had equal sensitivity for incident HIV cases among Black and White patients, while most simpler models had lower sensitivity for Black than White patients.

The full model identified 59·1% (378/640) of incident PrEP users among patients with high or very high risk scores, while simpler models identified fewer PrEP users. The distribution of HIV risk predictors was generally similar between patients with high or very high risk scores in the full model and patients actually prescribed PrEP, although fewer PrEP users were Black (3·6% vs. 28·4%; Appendix page 5).

DISCUSSION

Using predictive modeling with EHR data from 3·7 million members of a large healthcare system in California, we identified patients who were at high risk of HIV acquisition but not yet using PrEP. The full model had excellent discrimination between HIV cases and non-cases, with a C-statistic of 0·84 in the validation dataset. In contrast, existing HIV risk prediction tools have reported C-statistics of 0·66 to 0·72,[3] and prediction models commonly used in other areas of medicine have reported C-statistics of 0·71 to 0·82.[16] By flagging two percent of the general patient population as potential PrEP candidates, the full model identified nearly half of male HIV cases, and sensitivity for identifying an appropriate PrEP candidate is likely to be much higher. Although risk prediction tools are imperfect and cannot replace the clinical judgment of skilled providers, our model may be substantially more efficient than efforts to identify PrEP candidates in current practice. Our study suggests that HIV prediction models could be embedded in EHRs as an automated screening tool to help identify the subset of patients most likely to benefit from discussions about PrEP.

This study contributes to a growing body of literature on the use of EHR-based risk prediction tools to improve evidence-based preventive care. For example, automated cardiovascular risk prediction tools have been routinely embedded into EHRs to guide clinicians in appropriate aspirin and statin prescribing.[1721]In the HIV context, Krakower et al. evaluated multiple machine learning algorithms to predict incident HIV diagnosis using EHR data from an ambulatory practice in Massachusetts with 720,000 patients, with the best-performing model having comparable predictive performance to our full model.[18] Feller et al. used EHR data to predict incident HIV diagnoses in an academic medical center in New York City,[19] and Ridgway et al. used EHR data from an urban emergency department in Chicago to predict which patients met CDC criteria for PrEP use, subsequently using their model to prompt HIV risk assessments in clinical practice.[20] Our work builds on prior studies by including a large sample from a highly generalizable population with high-quality ascertainment of incident HIV diagnoses using registry data.

Our findings demonstrate the added value of rich EHR data for identification of potential PrEP candidates. Specifically, our results suggest that identification of potential PrEP candidates based on only MSM status or STI positivity, including our own efforts,[21] could be improved by incorporating other demographic and clinical data. Indeed, our weakest model included only STI positivity, the only CDC indication for PrEP use that is readily identifiable in EHR data.[1] Although collection of dozens of EHR variables may not be feasible for all healthcare systems that seek to identify PrEP candidates, we found that simpler models with only one to six variables provided some efficiencies in identifying patients at risk of incident HIV diagnosis, and had higher positive predictive value and specificity than the full model. Notably, the predictive performance of the model including only MSM status improved in recent years (data not shown), which may be attributable to efforts to improve collection of data on sexual orientation in EHRs; completeness of sexual orientation data increased from 16% in 2007-2014 to 40% in 2015-2017 at KPNC.

Our models did not identify cases among females, whose HIV risk may be largely dependent on the risk factors of their partners. Prior studies have used data collected in clinical trials and prospective cohorts to develop HIV risk prediction tools for African women and heterosexual couples,[22,23] but such tools have not been developed for these populations in settings with lower HIV incidence.[3] Inclusion of additional EHR variables, such as history of pelvic inflammatory disease or intimate partner violence, may improve identification of females at risk of HIV acquisition. Prediction might also improve with sex-stratified models, but we did not have sufficient HIV cases for a female-only model; a separate model for males performed similarly to the full model (data not shown). Nevertheless, given small numbers of transwomen and the difficulty of identifying HIV risk predictors in ciswomen, it may remain challenging to develop prediction models for females in the U.S.

Prior studies have found that existing HIV risk prediction tools based on CDC criteria for PrEP use underestimate HIV risk among Black MSM,[8] and there is mounting evidence that machine learning algorithms for risk prediction can be inadvertently racially biased.[24] We found that our full model had equal sensitivity for incident HIV cases among Black and White patients, while simpler models generally had lower sensitivity for Black compared with White patients. Traditional HIV risk factors such as MSM status and STI positivity may be less prevalent in the EHRs of Black individuals because of medical mistrust among patients,[25] poor communication between patients and providers,[26] or structural bias in the healthcare system.[27] Our results suggest that inclusion of other EHR data that do not rely on patient or provider behavior, such as location of residence, may reduce racial bias in HIV risk prediction tools based on EHR data. Notably, 28% of patients identified as high risk by the full model were Black, compared with only 4% of PrEP users, suggesting that implementing EHR-based prediction tools could help mitigate racial disparities in PrEP uptake.[2]

There were several limitations to this study. First, we did not externally validate our model in a different clinical setting. However, we validated our model in the most recent years of data, with results suggesting the full model would perform well if implemented prospectively in our large healthcare system. Second, there was potential for misclassification in our data, such as misclassification of HIV status for patients who were infected but not yet diagnosed, or PrEP status for patients who were prescribed PrEP outside of KPNC. Third, there were too few incident HIV diagnoses among transgender patients to assess sensitivity by gender identity. Fourth, to ensure sufficient baseline data to predict HIV risk, we restricted our study population to patients with at least two years of prior health plan enrollment and an outpatient visit at KPNC. Thus, our models may not perform as well among patients who have a shorter duration of enrollment or access medical care less regularly. Fifth, many patients in the 2015-2017 dataset were followed for less than three years; as a result, for patients with very high risk scores, the observed three-year risk of HIV was lower than that predicted by the full model. An additional limitation of our modeling approach was that we did not account for incident PrEP use as a competing risk for HIV acquisition; however, PrEP use was relatively rare during the time period for model development (2007-2014), and our model performed similarly in the validation dataset. Finally, our study population did not include adolescents, potentially limiting generalizability to this population.

Our study also had several strengths. First, our model development and validation processes were strengthened by the relatively large number of incident HIV cases, which allowed us to assess predictive performance by sex and race. Second, we compared the performance of a full model with that of simpler models, providing insight into the predictive value of additional EHR data domains for identifying potential PrEP candidates. Third, the KPNC HIV registry allowed for high-quality ascertainment of incident HIV diagnoses. Finally, the KPNC membership mirrors the age, sex, and race/ethnicity distributions of the surrounding population,[10] and the demographics of HIV-infected members are comparable to those of reported AIDS cases in California, strengthening the generalizability of our findings to the broader insured population.

In summary, EHR-based prediction models can identify patients at high risk of HIV acquisition who are not yet using PrEP. Models that include only MSM status and STI positivity do not perform as well as those including other data domains, particularly among Black individuals; nevertheless, the addition of only a handful of key variables can improve identification of patients who may benefit from PrEP. Although the risk predictors in our models may not generalize to all healthcare systems, our approach could be replicated in any clinical setting with an EHR. Additional studies should optimize EHR-based HIV risk prediction tools for females and for use in higher-incidence settings, such as safety net clinics, and evaluate their impact on PrEP prescribing and HIV incidence.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

We thank Susan Gruber, PhD and Jenna Wong, PhD for helpful advice on our analytic approach. This work was supported by the Kaiser Permanente Northern California Community Benefit Research Program, the National Institute of Allergy and Infectious Diseases (K01 AI122853 to JLM), and the National Institute of Mental Health (K23 MH098795 to DSK).

REFERENCES

1. Preexposure prophylaxis for the prevention of HIV infection in the United States -- 2017 update. Centers for Disease Control and Prevention 2017 https://www.cdc.gov/hiv/pdf/guidelines/cdc-hiv-PrEPguidelines-2017.pdf.

2. Huang YA, Zhu W, Smith DK, Harris N, Hoover KW. HIV Preexposure Prophylaxis, by Race and Ethnicity - United States, 2014-2016. MMWR Morb Mortal Wkly Rep 2018; 67(41): 1147–50. [PubMed: 30335734]

3. Chou R, Evans C, Hoverman A, et al. Pre-Exposure Prophylaxis for the Prevention of HIV Infection: A Systematic Review for the U.S. Preventive Services Task Force. Evidence Synthesis No. 178. AHRQ Publication No. 18-05247-EF-1. Rockville, MD: Agency for Healthcare Research and Quality, 2018.

4. Silapaswan A, Krakower D, Mayer KH. Pre-Exposure Prophylaxis: A Narrative Review of Provider Behavior and Interventions to Increase PrEP Implementation in Primary Care. J Gen Intern Med 2017; 32(2): 192–8. [PubMed: 27761767]

5. Burke RC, Sepkowitz KA, Bernstein KT, et al. Why don't physicians test for HIV? A review of the US literature. AIDS 2007; 21(12): 1617–24. [PubMed: 17630557]

6. Smith DK, Pals SL, Herbst JH, Shinde S, Carey JW. Development of a clinical screening index predictive of incident HIV infection among men who have sex with men in the United States. J Acquir Immune Defic Syndr 2012; 60(4): 421–7. [PubMed: 22487585]

7. Haukoos JS, Lyons MS, Lindsell CJ, et al. Derivation and validation of the Denver Human Immunodeficiency Virus (HIV) risk score for targeted HIV screening. Am J Epidemiol 2012; 175(8): 838–46. [PubMed: 22431561]

8. Lancki N, Almirol E, Alon L, McNulty M, Schneider JA. Preexposure prophylaxis guidelines have low sensitivity for identifying seroconverters in a sample of young Black MSM in Chicago. AIDS 2018; 32(3): 383–92. [PubMed: 29194116]

9. Jones J, Hoenigl M, Siegler AJ, Sullivan PS, Little S, Rosenberg E. Assessing the Performance of 3 Human Immunodeficiency Virus Incidence Risk Scores in a Cohort of Black and White Men Who Have Sex With Men in the South. Sex Transm Dis 2017; 44(5): 297–302. [PubMed: 28407646]

10. Gordon N. Similarity of the Adult Kaiser Permanente Membership in Northern California to the Insured and General Population in Northern California: Statistics from the 2011 California Health Interview Survey. 2015 https://divisionofresearch.kaiserpermanente.org/projects/memberhealthsurvey/SiteCollectionDocuments/chis_non_kp_2011.pdf (accessed January 17, 2019.

11. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD). Ann Intern Med 2015; 162(10): 735–6.

12. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition. New York: Springer-Verlag; 2009.

13. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010; 33(1): 1–22. [PubMed: 20808728]

14. Robins JM, Finkelstein DM. Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. Biometrics 2000; 56(3): 779–88. [PubMed: 10985216]

15. Pencina MJ, D'Agostino RB Sr. Evaluating Discrimination of Risk Prediction Models: The C Statistic. JAMA 2015; 314(10): 1063–4. [PubMed: 26348755]

16. Goff DC Jr., Lloyd-Jones DM, Bennett G, et al. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. Circulation 2014; 129(25 Suppl 2): S49–73. [PubMed: 24222018]

17. Sekaran NK, Sussman JB, Xu A, Hayward RA. Providing clinicians with a patient’s 10-year cardiovascular risk improves their statin prescribing: a true experiment using clinical vignettes. BMC Cardiovasc Disord 2013; 13: 90. [PubMed: 24148829]

18. Krakower D, Gruber S, Menchaca JT, et al. Automated identification of potential candidates for HIV preexposure prophylaxis using electronic health record data IDWeek. New Orleans, Louisiana; 2016.

19. 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. [PubMed: 29084046]

20. Ridgway JP, Almirol EA, Bender A, et al. Which Patients in the Emergency Department Should Receive Preexposure Prophylaxis? Implementation of a Predictive Analytics Approach. AIDS Patient Care STDS 2018; 32(5): 202–7. [PubMed: 29672136]

21. Guinness RR, Volk JE, Hurley LB, Tobias TT, Marcus JL. Low-Intensity Outreach to Increase Uptake of HIV Preexposure Prophylaxis Among Patients with Sexually Transmitted Infections. AIDS Behav 2019; 23(2): 544–7. [PubMed: 30101394]

22. Irungu EM, Heffron R, Mugo N, et al. Use of a risk scoring tool to identify higher-risk HIV-1 serodiscordant couples for an antiretroviral-based HIV-1 prevention intervention. BMC Infect Dis 2016; 16(1): 571. [PubMed: 27751179]

23. Balkus JE, Brown E, Palanee T, et al. An Empiric HIV Risk Scoring Tool to Predict HIV-1 Acquisition in African Women. J Acquir Immune Defic Syndr 2016; 72(3): 333–43. [PubMed: 26918545]

24. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. N Engl J Med 2018; 378(11): 981–3. [PubMed: 29539284]

25. Eaton LA, Driffin DD, Kegler C, 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.

26. Ashton CM, Haidet P, Paterniti DA, et al. Racial and ethnic disparities in the use of health services: bias, preferences, or poor communication? J Gen Intern Med 2003; 18(2): 146–52. [PubMed: 12542590]

27. Nelson A. Unequal treatment: confronting racial and ethnic disparities in health care. J Natl Med Assoc 2002; 94(8): 666–8. [PubMed: 12152921]

RESEARCH IN CONTEXT

Evidence before this study

Antiretroviral preexposure prophylaxis (PrEP) is more than 90% effective in preventing HIV infection but uptake has been limited, in part because providers face challenges identifying patients who may benefit from PrEP. We searched PubMed for studies published from Jan 1, 2007, to December 1, 2018, with the terms “HIV” and “risk prediction,” “risk score,” or “risk index.” Existing HIV prediction tools require that providers already know a patient belongs to a given risk group, such as men who have sex with men (MSM), and require additional collection of data from patients on HIV risk behaviors, limiting their use in clinical practice. Moreover, such tools underestimate risk for HIV acquisition in Black MSM.

Added value of this study

Using EHR data from 3·7 million members of a large healthcare system in California, this study demonstrated that predictive modeling can be used to identify patients who are at high risk of HIV acquisition in a general patient population. By flagging two percent of the general patient population as potential PrEP candidates, the model identified nearly half of male HIV cases. This study also demonstrated the added value of rich EHR data for identification of potential PrEP candidates, with a full model outperforming simpler models based on only MSM status and recent bacterial sexually transmitted infections.

Implications of all the available evidence

HIV prediction models could be embedded in EHRs as an automated screening tool to help identify the subset of patients most likely to benefit from discussions about PrEP. Future studies should optimize EHR-based HIV risk prediction tools and evaluate their impact on PrEP prescribing and HIV incidence.

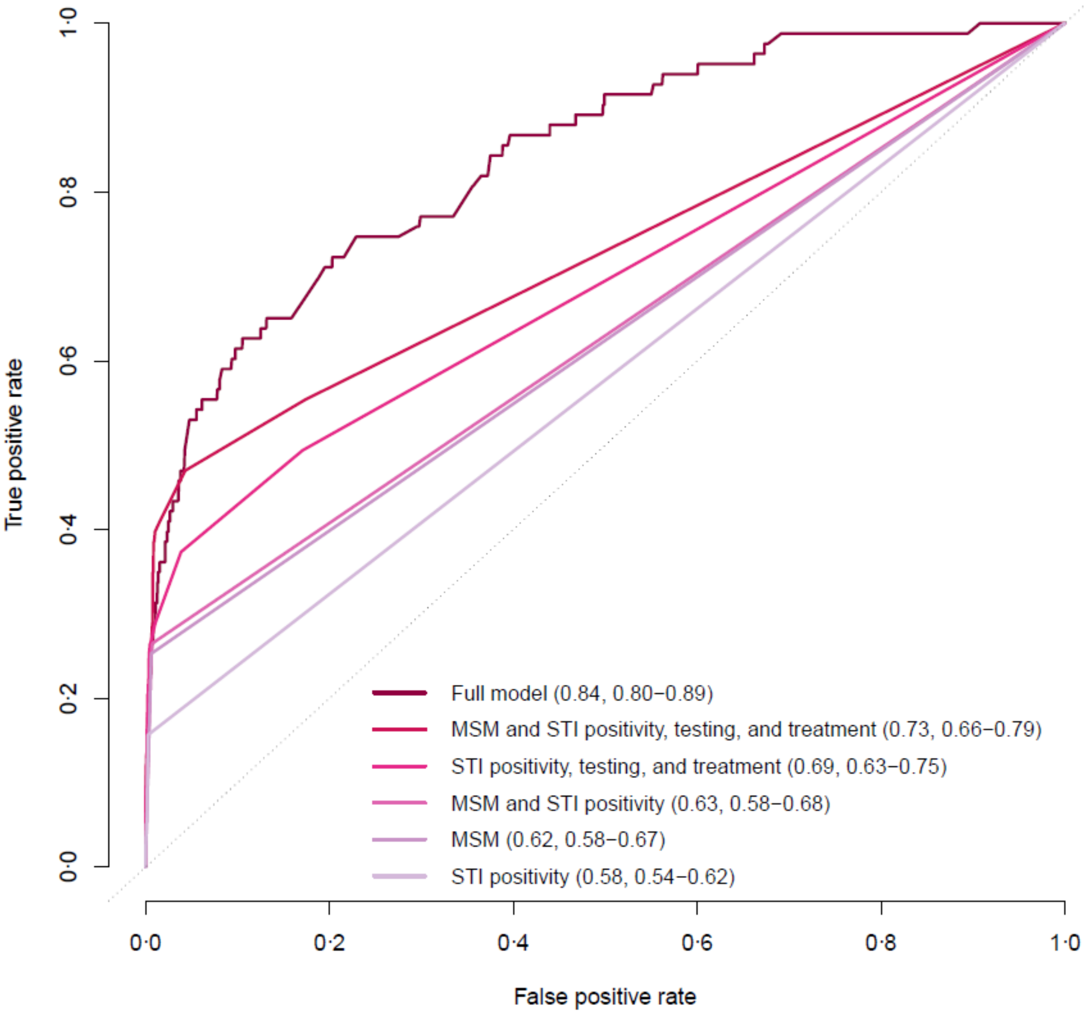

Figure 1. Receiver operating characteristic curves and C-statistics for full and simpler HIV risk prediction models in the validation dataset – Kaiser Permanente Northern California, 2015-2017 (N=606,701).

STI, sexually transmitted infection; MSM, men who have sex with men. C-statistics were computed as area under the curve and reported in parentheses, followed by 95% confidence intervals obtained by bootstrapping with 1000 resamples of the data.

Table 1.

Demographic characteristics and incident HIV diagnoses within three years among patients in the development and validation datasets – Kaiser Permanente Northern California, 2007-2017 (N=3,750,664)

Characteristic	Development dataset, 2007-2014 (n=3,143,963)	Validation dataset, 2015-2017 (n=606,701)
Male, n (%)	1,462,453 (46·5)	297.024 (49·0)
Sexual orientation, n (% among known)[a]		
Heterosexual	472,324 (96·4)	233,365 (95·5)
Gay or lesbian	14,083 (2·9)	8314 (3·4)
Bisexual	3584 (0·7)	2641 (1·1)
Unknown sexual orientation, n (%)	2,653,972 (84·4)	362,381 (59·7)
Age, mean (SD)	44·6 (17·9)	37·4 (16·2)
Race/ethnicity, n (% among known)		
White	1,521,081 (51·9)	251,520 (44·0)
Hispanic	565,576 (19·3)	138,861 (24·3)
Asian	504,896 (17·2)	131,410 (23·0)
Black	217,561 (7·4)	36,589 (6·4)
Other	120,775 (4·1)	13,415 (2·3)
Unknown race/ethnicity, n (%)	214,074 (6·8)	34,906 (5·8)
Received care in 1 of 3 cities with high HIV incidence, n (%)	438,462 (13·9)	91,953 (15·2)
Resided in 1 of 8 urban ZIP codes with high HIV incidence, n (%)	70,105 (2·2)	16,779 (2·8)
Incident HIV diagnosis within 3 years, n (%)	701 (<0·1)	83 (<0·1)

[a] Sexual orientation data was based on sex, and gender of sex partners, as recorded in the electronic health record.

Table 2.

HIV risk predictors retained in the full LASSO model, prevalence among patients with and without an incident HIV diagnosis within three years, and adjusted odds ratios in the development dataset – Kaiser Permanente Northern California, 2007-2014 (N=3,143,963)

	Predictor prevalence		
HIV risk predictor variable	**No incident HIV (n=3,143,262)**	**Incident HIV (n=701)**	**Adjusted odds ratio**
Demographics and social history			
Male, n (%)	1,461,821 (46·5)	632 (90·2)	7·2
Men who have sex with men, n (%)	5209 (0·2)	41 (5·8)	4·7
Sexually active, n (%)	534,303 (17·0)	77 (11·0)	0·8
Aged 50-59, n (%)	499,800 (15·9)	88 (12·6)	0·8
Aged ≥60, n (%)	647,327 (20·6)	29 (4·1)	0·4
Black, n (%)	217,418 (6·9)	143 (20·4)	2·6
Hispanic, n (%)	565,435 (18·0)	141 (20·1)	1·1
Asian, n (%)	504,823 (16·1)	73 (10·4)	0·9
Other race/ethnicity, n (%)	120,739 (3·8)	36 (5·1)	1·1
Neighborhood deprivation index quintile 2, n (%)[a]	631,340 (20·1)	113 (16·1)	0·99
Neighborhood deprivation index quintile 3, n (%)	633,170 (20·1)	98 (14·0)	0·96
Neighborhood deprivation index quintile 4, n (%)	467,367 (14·9)	126 (18·0)	1·03
Received care in 1 of 3 cities with high HIV incidence, n (%)[b]	438,205 (13·9)	257 (36·7)	1·1-2·4
Resided in 1 of 8 urban ZIP codes with high HIV incidence, n (%)	1,461,821 (46·5)	632 (90·2)	1·1-3·0
Laboratory tests and results			
Positive urine test for methadone, ever prior, n (%)	264 (<0·1)	1 (0·1)	13·9
Positive urine test for cocaine, ever prior, n (%)	1745 (0·1)	2 (0·3)	1·1
Number of HIV testing episodes, prior 2 years, mean (SD)	0·14 (0·43)	0·59 (1·07)	1·1
Number of HIV antibody or RNA tests, prior 2 years, mean (SD)	0·14 (0·43)	0·59 (1·09)	1·3
Number of tests for rectal gonorrhea or chlamydia, ever prior, mean (SD)	0 (0·03)	0·03 (0·22)	0·8
Number of positive tests for rectal gonorrhea or chlamydia, prior 2 years, mean (SD)	0 (0·01)	0 (0·08)	2·0
Number of positive tests for urethral chlamydia, prior 2 years, mean (SD)	0 (0·08)	0·03 (0·18)	0·9
Number of positive tests for urethral gonorrhea, prior 2 years, mean (SD)	0 (0·03)	0·04 (0·21)	3·5
Number of RPR or treponemal tests for syphilis, prior 2 years, mean (SD)	0·15 (0·46)	0·49 (0·86)	1·4

HIV risk predictor variable	Predictor prevalence No incident HIV (n=3,143,262)	 Incident HIV (n=701)	Adjusted odds ratio
Number of reactive RPR or positive treponemal tests for syphilis, prior 2 years, mean (SD)	0 (0·04)	0·04 (0·24)	1·1
Medication use			
Medications for erectile dysfunction, ever prior, n (%)	132,914 (4·2)	71 (10·1)	1·02
Number of penicillin G benzathine injections with syphilis test within 90 days, prior 2 years, mean (SD)	0 (0·01)	0·01 (0·12)	4·3
Diagnoses			
Number of anal wart diagnoses, ever prior, mean (SD)	0 (0·04)	0·03 (0·42)	1·5
Depression, ever prior, n (%)	91542 (2·9)	33 (4·7)	1·2
Any psychiatric diagnosis, ever prior, n (%)[c]	135317 (4·3)	49 (7·0)	1·002
Transgender-related diagnosis, ever prior, n (%)	911 (<0·1)	3 (0·4)	2·0
High-risk sexual behavior (homosexual), ever prior, n (%)	27 (<0·1)	0 (0)	0·7
High-risk sexual behavior (not specified), ever prior, n (%)	453 (<0·1)	3 (0·4)	2·6
Exposure to HIV, ever prior, n (%)	304 (<0·1)	4 (0·6)	1·1
HIV counseling, ever prior, n (%)	2858 (0·1)	7 (1·0)	1·01
HIV education, ever prior, n (%)	6663 (0·2)	15 (2·1)	1·1

LASSO, least absolute shrinkage and selection operator; SD, standard deviation; RPR, rapid plasma reagin. Table includes all 44 predictors retained in the full model by the LASSO variable selection procedure. Denominators were the full sample, not only those with known data.

[a] The highest quintile of the neighborhood deprivation index (5) indicates the greatest neighborhood deprivation, and the lowest quintile (1) indicates the least neighborhood deprivation.[12]

[b] High-incidence cities and ZIP codes were 3 and 8 separate variables in the model, respectively.

[c] Any psychiatric diagnosis included depression, bipolar disorder, schizophrenia, and attention deficit disorder.

Table 3.

Performance of full and simpler HIV risk prediction models among patients with high or very high risk scores in the validation dataset – Kaiser Permanente Northern California, 2015-2017 (N=606,701)

		Sensitivity (% of incident HIV cases identified)								
HIV risk prediction model	**N of patients flagged**	**Overall (n=83)**	**Male (n=69)**	**Female (n=14)**	**Black (n=20)**	**White (n=20)**	**Sensitivity for incident PrEP users, % (n=640)**	**Specificity, %**	**PPV, %**	**NPV, %**
Full LASSO model	13,463	38·6	46·4	0	25·0	25·0	59·1	97·8	0·24	>99·9
MSM status and STI positivity, testing, and treatment	4617	28·9	34·8	0	15·0	35·0	42·7	99·2	0·52	>99·9
STI positivity, testing, and treatment	1458	20·5	24·6	0	10·0	15·0	16·2	99·8	1·17	>99·9
MSM status and STI positivity	3849	25·3	30·4	0	10·0	30·0	38·8	99·4	0·55	>99·9
MSM status	3790	25·3	30·4	0	10·0	30·0	38·1	99·4	0·55	>99·9
STI positivity	1016	6·0	7·2	0	10·0	0	6·4	99·8	0·49	>99·9

LASSO, least absolute shrinkage and selection operator; STI, sexually transmitted infection; MSM, men who have sex with men; PrEP, preexposure prophylaxis; PPV, positive predictive value; NPV, negative predictive value. High risk scores were defined as a predicted probability of an incident HIV diagnosis within three years of 0·20-0·99%, and very high risk scores as a predicted probability of ≥1·0%.