Published in final edited form as:

Lancet HIV. 2019 October ; 6(10): e696–e704. doi:10.1016/S2352-3018(19)30139-0.

Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study

Douglas S. Krakower, MD[1,2,3,4], **Susan Gruber, PhD**[4], **Katherine Hsu, MD**[5,6], **John T. Menchaca, BA**[4], **Judith C. Maro, PhD**[4], **Benjamin A. Kruskal, MD**[7,8], **Ira B. Wilson, MD**[9], **Kenneth H. Mayer, MD**[1,2,3], **Michael Klompas, MD**[3,4,10]

[1]Division of Infectious Diseases, Beth Israel Deaconess Medical Center, Boston, MA

[2]The Fenway Institute, Fenway Health, Boston, MA

[3]Harvard Medical School, Boston, MA

[4]Department of Population Medicine, Harvard Medical School, Boston, MA

[5]Bureau of Infectious Disease and Laboratory Sciences, Massachusetts Department of Public Health, Boston, MA

[6]Department of Pediatrics, Boston Medical Center, Boston, MA

[7]Atrius Health, Boston, MA

[8]New England Quality Care Alliance, Braintree, MA

[9]Department of Health Services, Policy and Practice, Brown University, Providence, RI

[10]Division of Infectious Diseases, Brigham and Women's Hospital, Boston, MA

Abstract

Contact information for the corresponding author: Douglas S. Krakower, MD, Division of Infectious Diseases, Beth Israel Deaconess Medical Center, 110 Francis St., Lowry Medical Office Building, Suite GB, Boston, MA, 02215, dkrakowe@bidmc.harvard.edu, Phone: (617) 632-0758; Fax: (617) 632-7626.

AUTHOR CONTRIBUTIONS

Concept and design: Krakower, Klompas, Hsu, Mayer, Wilson, Gruber, Maro, Kruskal.
Acquisition, analysis, or interpretation of data: Klompas, Krakower, Gruber, Hsu, Menchaca.
Drafting of the manuscript: Krakower, Klompas.
Critical revision of the manuscript for important intellectual content: All authors.
Statistical analysis: Gruber.
Obtained funding: Krakower, Klompas, Hsu, Maro, Mayer, Wilson.
Administrative, technical, or material support: Klompas, Krakower.
Supervision: Krakower, Klompas, Hsu.
Author access to data: Drs. Krakower and Klompas had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

DECLARATION OF INTERESTS

DK has conducted research supported by Gilead Sciences and has received funding to author continuing medical education content for Medscape, MED-IQ, DKBmed, and UptoDate, Inc. KHM has conducted research supported by Gilead Sciences and ViiV and has received funding to author continuing medical education content for UptoDate, Inc. All authors declare that there are no other potential conflicts of interest.

Background: HIV preexposure prophylaxis (PrEP) is effective but underutilized, in part because clinicians lack tools to identify PrEP candidates. We developed and validated an automated prediction algorithm using electronic health records (EHR) data to identify individuals at increased risk for HIV acquisition.

Methods: We used machine learning algorithms to predict incident HIV infections using 180 potential predictors of HIV risk drawn from EHR data from 2007-2015 at Atrius Health, an ambulatory group practice in Massachusetts, USA. The best-performing model was validated prospectively using 2016 data from Atrius Health and externally using 2011-2016 data from Fenway Health, a community health center specializing in sexual healthcare in Boston, Massachusetts. We assessed the model's performance at identifying individuals with incident HIV and patients independently prescribed PrEP by clinicians using cross-validated area under the curve (cv-AUC).

Findings: Cohorts included 1,155,966 Atrius Health patients from 2007-2015 (including 150 [<0·1%] patients with incident HIV), 537,257 patients in 2016 (16 [<0·1%] with incident HIV), and 33,404 Fenway Health patients from 2011-2016 (423 [1·3%] with incident HIV). The best-performing algorithm had a cv-AUC of 0·86 (95% CI 0·82-0·90) for identifying incident HIV infections in the development cohort, 0·91 (95% CI 0·81-1·00) on prospective validation, and 0·77 (95% CI 0·74-0·79) on external validation. The model successfully identified patients independently prescribed PrEP by clinicians at Atrius Health (cv-AUC 0·94, 95% CI 0·90-0·97) or Fenway Health (cv-AUC 0·79, 95% CI 0·78-0·80). HIV risk scores increased steeply at the 98th percentile. We designated patients with scores above this threshold as potential PrEP candidates and prospectively identified 9,515/536,384 (1·8%) new PrEP candidates at Atrius Health in 2016.

Interpretation: Automated algorithms can efficiently identify patients at increased risk for HIV acquisition. Integrating these models into EHRs to alert providers about patients who may benefit from PrEP could improve PrEP prescribing and prevent new HIV infections.

Funding: The Harvard University Center for AIDS Research, the Providence/Boston Center for AIDS Research, the Rhode Island IDeA-CTR [U54GM11567], and the US Centers for Disease Control and Prevention.

INTRODUCTION

HIV preexposure prophylaxis (PrEP) decreases HIV incidence in high-risk populations.[1-4] In 2014, the US Centers for Disease Control and Prevention (CDC) recommended PrEP as an HIV prevention option for individuals at substantial risk for HIV infection.[5] To date, however, PrEP has been underutilized. The CDC estimates that 1·1 million Americans have indications for PrEP,[6] but only 100,000 individuals were prescribed PrEP in 2017.[7] Few primary care providers have prescribed PrEP.[8] Reasons for the low rate of prescribing include insufficient time to assess HIV risk during clinical visits, limited knowledge about PrEP, and uncertainty about whether providing PrEP lies within their clinical purview.[9] There is consequently a need for tools to help providers identify persons at high risk of HIV acquisition who may benefit from PrEP.[10]

Electronic clinical decision support using data embedded in patients' electronic health records (EHRs) could address this need and might empower more primary care providers to prescribe PrEP. Prior studies have demonstrated the utility of EHR data to predict important clinical outcomes[11,12] and inform clinical decision support interventions[13,14] in other areas of medicine. Potential predictors of HIV risk available in EHR data include demographics, sexually transmitted infection (STI) testing and diagnoses, diagnoses for viral hepatitis and substance use disorders, and suggestive prescriptions (e.g. empirical treatment for STIs, medications for opiate use disorders).

We developed prediction models for incident HIV using EHR data from a general primary care population. We envisioned a prediction model for incident HIV as a first-stage screening tool to prompt primary care providers to discuss interest and suitability for PrEP with higher risk patients. We derived the prediction model using data from Atrius Health, a healthcare organization serving a general primary care population, and validated it prospectively using an additional year of data from Atrius. We further validated the model by applying it to EHR data from Fenway Health, a community health center specializing in healthcare for sexual and gender minorities, and compared model predictions to incident HIV cases at Fenway Health and to clinicians' independent PrEP prescribing at both Atrius Health and Fenway Health.

METHODS

Study Setting and Design

Atrius Health provides ambulatory care at 32 clinical sites to 550,000 patients annually in Massachusetts, USA and uses an EHR system (EPIC, Verona, WI) for documenting clinical healthcare data. Fenway Health in Boston, Massachusetts provides primary care to 18,000 patients annually, and its primary care clinicians prescribe PrEP routinely. Fenway Health also uses an EHR system (Centricity™, Boston, MA) for documenting healthcare.

We derived a predictive model for incident HIV infection using Atrius Health data from Jan 1, 2007 to Dec 31, 2015 (development cohort) and validated its performance prospectively using Atrius Health data from Jan 1, 2016 to Dec 31, 2016 (prospective validation cohort), and externally using Fenway Health data from Jan 1, 2011 to Dec 31, 2016 (external validation cohort). We developed our model in accordance with the TRIPOD statement for prediction models.[15]

The primary outcome was incident HIV infection, defined as 1) an incident positive HIV enzyme-linked immunosorbent assay (ELISA) with confirmatory Western Blot with no prior evidence of positive HIV tests or HIV-related prescriptions; or 2) a positive HIV ELISA or Antibody/Antigen test following a negative HIV ELISA within the preceding 2 years and no prior evidence of positive HIV tests or HIV-related prescriptions. We also compared model predictions of HIV risk to clinicians' assessment of HIV risk as reflected by their independent PrEP prescribing. PrEP prescribing was defined as ≥2 prescriptions for tenofovir disoproxil fumarate with emtricitabine ≥2 months apart to a patient without evidence of HIV or chronic hepatitis B infection.

Predictors

Multiple categories of EHR data were extracted to maximize the breadth of potential predictor variables, including demographics, diagnoses, prescriptions, and laboratory tests. Expert physicians proposed 180 EHR variables potentially associated with HIV risk. Of these, 46 variables were removed because their values were zero for all patients or because they were identical to another variable in the dataset, leaving 134 covariates for model development (appendix p 1). EHR data on gender identity and gender of sexual partners were not available at Atrius Health. To improve identification of men who have sex with men and transgender persons, we included variables suggestive of anal sex, such as testing for rectal STIs and diagnosis codes for anal dysplasia, and diagnosis codes associated with transgender status.

EHR data were extracted using ESP (Electronic medical record Support for Public Health, esphealth.org), a generalizable open-source public health surveillance platform for analyzing and communicating EHR data to public health departments, practice managers, or clinicians as appropriate.[16]

Study Cohorts

All Atrius Health patients aged ≥15 with at least 1 clinical encounter during 2007-2015 were queried by ESP for any of the following: 1) HIV infection; 2) PrEP prescriptions; or 3) any of the potential EHR predictor variables (appendix p 1). The prospective validation cohort comprised patients aged ≥15 seen during 2016. The external validation cohort included all Fenway Health patients seen during 2011-2016.

Model Development

We developed 42 candidate prediction models using machine learning and logistic regression models and compared their performances at discriminating between patients with incident HIV and matched control patients. Cases of incident HIV were matched by sex to up to 50 controls. We developed and compared candidate models using Super Learning (appendix pp 5,6).

Model Validation

Our primary measure of algorithm performance was discrimination, the ability to separate individuals who developed HIV from those who did not. We focused on discrimination because our goal was to stratify patients according to their risk of incident HIV to identify potential PrEP candidates. We measured discrimination using ten-fold cross-validated area under the receiver-operating curve (cv-AUC) weighted to account for case-control sampling. After identifying the candidate algorithm with the highest cv-AUC using Atrius Health data from 2007-2015, we validated its performance prospectively using 2016 data. To assess generalizability to a population with high HIV incidence, we measured this model’s cv-AUC for incident HIV cases at Fenway Health. We also calculated the cv-AUC for identifying patients receiving PrEP prescriptions at Atrius Health and Fenway Health to compare model predictions to providers’ independent clinical decisions. Model calibration, the agreement between observed outcomes and predictions, was assessed by comparing HIV incidence at Atrius Health in 2016 to model predictions.

Identification of PrEP Candidates

We identified potential PrEP candidates at Atrius Health by calculating HIV risk scores (i.e. probability of an incident HIV diagnosis) for every HIV-uninfected patient not on PrEP during 2007-2015. We then inspected the distribution of scores for an inflection point to define a subgroup of patients with elevated scores relative to the general population who might thus represent possible candidates for PrEP.

We calculated numbers of patients with incident HIV or PrEP prescriptions who had scores above the inflection point and selected alternative thresholds, such as the 90th percentile score for the general population, the median score for patients with incident HIV, and the median score for patients independently prescribed PrEP by clinicians. We calculated sensitivity, specificity, and positive and negative predictive values for all thresholds.

Study procedures were approved by the IRB at Harvard Pilgrim Health Care with a waiver of written informed consent.

Role of the Funding Sources

The study sponsors had no role in study design; collection, analysis, and interpretation of data; writing of the report; or the decision to submit the paper for publication.

RESULTS

Of 1,155,966 Atrius Health patients seen during 2007-2015, 150 (<0·1%) were diagnosed with incident HIV and 90 (<0·1%) initiated PrEP. There were 537,257 patients seen at Atrius Health in 2016 (16 [<0·1%] with incident HIV, 128 [<0·1%] initiating PrEP), and 33,404 at Fenway Health during 2011-2016 (423 [1·3%] with incident HIV, 1,813 [5·4%] initiating PrEP) (Table 1). The proportion of Fenway Health patients with incident HIV (1,300 per 100,000) was 100 times greater than in the Atrius Health population (13 per 100,000). The cohorts were similar with respect to racial and ethnic composition, but Fenway Health patients included 6·5% transgender or gender non-conforming patients while gender at Atrius Health was only recorded as binary.

The majority of incident HIV infections and nearly all PrEP prescriptions were in men in both practices (appendix p 7). Patients with incident HIV were disproportionately Black and those prescribed PrEP were disproportionately White at both healthcare organizations.

Using the development cohort, weighted cv-AUCs for the 42 candidate prediction algorithms ranged from 0·42 to 0·86. The highest cv-AUC was obtained using LASSO (least absolute shrinkage and selection operator) (cv-AUC 0·86, 95% CI 0·82-0·90, Figure 1). LASSO's automated variable selection procedure retained 23 predictor variables in the final model (Table 2). These included diagnosis codes (e.g. syphilis and HIV counseling), laboratory tests (e.g. numbers of HIV tests), prescriptions (e.g. penicillin G benzathine), and registration data (e.g. race). This model was used to generate all subsequently reported prediction scores.

The LASSO model had good prospective discrimination, with a cv-AUC of 0·91 (95% CI 0·81-1·00) using 2016 data. Discrimination was also good when the model was applied externally to Fenway Health (cv-AUC 0·77, 95% CI 0·74-0·79), and when applied to detect PrEP patients at Atrius Health (cv-AUC 0·94, 95% CI 0·90-0·97) and Fenway Health (cv-AUC 0·79, 95% CI 0·78-0·80).

Of 537,257 Atrius Health patients seen in 2016, 16 (<0.1%) had incident HIV. This proportion of new cases (3·0 per 100,000) is larger than the 2·3 per 100,000 expected with our model. Upon further inspection, 3 of the incident HIV cases were first seen at Atrius Health in 2016 and thus had no historical data to inform their risk prediction. When excluding these cases, the proportion of remaining cases with incident HIV (2·4 per 100,000) indicates the LASSO model was well-calibrated among patients with at least 1 year of historic EHR data.

We used the LASSO model to calculate HIV risk scores for all Atrius Health patients seen during 2007-2015; patients with no recorded risk factors were assigned scores of 0. Predicted risk ranged from 0 to 95,000 out of 100,000, with a median of 0 (interquartile range [IQR] 0 to 1·6 out of 100,000), and a mean of 22 out of 100,000. A marked increase in risk scores was seen at the 98th percentile (Table 3). Amongst 1,154,724 Atrius Health patients without HIV or PrEP use during 2007-2015, 23,018 (2·0%) had scores above the 98th percentile and were defined as potential PrEP candidates, as were 9,515/536,384 (1·8%) Atrius Health patients seen in 2016 and 4,385/28,702 (15·3%) Fenway Health patients.

Six (37.5%) of the 16 patients with incident HIV and 62/128 (48·4%) patients initiating PrEP at Atrius Health in 2016 had risk scores above the 98th percentile and would thus have been flagged by the model as potential PrEP candidates using this threshold (Table 3). At Fenway Health, 196/423 (46·3%) patients with incident HIV and 851/1,813 (46·9%) PrEP patients would have been flagged. Table 3 illustrates how varying the threshold for flagging patients would affect the number of potential PrEP candidates in each study cohort. For example, at the 90th percentile, the model would identify 15/16 (94%) Atrius patients newly diagnosed with HIV in 2016 as PrEP candidates and 115/128 (89·8%) of patients prescribed PrEP, as well as 386/423 (91·3%) Fenway patients with incident HIV and 1,721/1,813 (94·9%) PrEP users. However, using this alternative threshold would also increase the number of patients flagged by the algorithm to 48,533/536,384 (9·0% of the population) at Atrius Health in 2016, and 16,023/28,702 (55·8% of the population) at Fenway Health.

Sensitivity and specificity of the model for detecting incident HIV varied for different risk score thresholds in the development cohort (Table 4) and the validation cohorts (appendix p 8). Positive predictive values were low and negative predictive values were high at all score thresholds for all study cohorts.

In the Atrius Health population during 2007-2015, HIV prediction scores for patients with incident HIV (median score 25 out of 100,000, IQR 8·6 to 76 out of 100,000) were higher than scores for the general population (median score 0, IQR 0 to 1·6 out of 100,000; p <0·0001), as were scores for patients using PrEP (median score 42 out of 100,000, IQR 29 to 250 out of 100,000; p <0·0001). There were 22,893/1,154,724 (2·0%) patients with scores above the median score of patients with incident HIV, and 6,647/1,154,724 (0·6%) with scores above the median score of patients prescribed PrEP.

DISCUSSION

We used EHR data from over 1 million patients to develop and validate an automated prediction model to identify patients at increased risk for HIV acquisition and therefore potential candidates for PrEP. Our model discriminated well between patients with and without incident HIV (cv-AUC 0·86-0·91) and between patients with and without PrEP prescriptions (cv-AUC 0·93) in a general primary care population. We identified many patients with HIV risk scores substantially higher than the general population who were not prescribed PrEP, therefore providing an opportunity to prompt providers to conduct targeted discussions with patients about their eligibility and interest in PrEP.

The HIV prediction model we developed was fitted using data from a general primary care organization but could also discriminate between patients with and without incident HIV in an independent practice serving sexual and gender minorities and among patients whom expert clinicians deemed suitable for PrEP. These findings indicate this model may be generalizable to diverse healthcare organizations.

Predictive performance was lower at Fenway Health, possibly because of differences in HIV epidemiology, patterns of healthcare, and EHR usage. Model performance for distinct populations could potentially be improved by model building directly upon data from specific target populations.

Many patients with incident HIV or PrEP use in our study had low HIV risk scores, illustrating that some patients at risk for HIV acquisition, or deemed suitable for PrEP by clinicians, are not identified as PrEP candidates using the score cut-offs we evaluated. Presumably these patients' HIV risk behaviors did not result in distinctive EHR profiles, or these were patients with little historical data. Possible explanations for non-suggestive EHR profiles include patients not offering and/or providers not eliciting information about HIV risk behaviors,[17,18] information on risk behaviors being recorded only in free text, failure of providers to document or act on disclosed risk behaviors (e.g. with appropriate STI testing), or receipt of pertinent healthcare externally (e.g. attending walk-in STI clinics). For PrEP, providers may also be prescribing to risk-averse patients who request PrEP despite their low risk for HIV.[9] Our model's inability to identify all patients at risk of incident HIV underscores the importance of integrating the results of prediction models with routine, comprehensive HIV risk assessments by knowledgeable clinicians.[19] Prediction models using historical EHR data might also overestimate some patients' current HIV risk. Thus, alerts about high-risk patients should prompt patient-provider discussions about PrEP and not necessarily result in prescriptions. Nonetheless, our models correctly identified 6 of the 16 (37·5%) patients diagnosed with incident HIV in 2016 while flagging only 2% of the general population. If clinicians had discussed PrEP with all patients with risk scores above our cut-off and prescribed PrEP to those who indicated current high-risk behaviors, our model could have helped avert nearly 40% of the new HIV infections at Atrius Health in 2016. We believe that this cut-off would identify a large proportion of high-risk patients without placing an undue burden on clinicians. Machine learning algorithms thus provide an efficient means to screen large populations for high-risk individuals who merit HIV testing and consideration for PrEP.

Disparities exist in PrEP uptake in the US, with racial and ethnic minorities underrepresented among PrEP users despite being at increased risk for HIV.[20] Lower PrEP uptake among minorities may result from structural challenges to accessing preventive health care, such as insurance and financial barriers, medical mistrust, and providers' implicit biases.[21] Machine learning algorithms could inadvertently exacerbate disparities in PrEP provision if based on health care variables that vary by race, ethnicity, gender, or other patient characteristics[22] (e.g. insurance status, income). Conversely, machine learning algorithms might also mitigate providers' biases in PrEP prescribing by providing objective risk assessments.[23] Our final model included Black race and primary language as predictors, suggesting our model could improve identification of minority individuals. Future implementation studies will need to examine how HIV prediction models impact disparities in PrEP use.

This study adds to the literature on machine learning algorithms to predict important clinical outcomes using EHR data. These algorithms have been used to predict nosocomial *Clostridioides difficile* infection, and clinical outcomes from sepsis, myocardial infarction, Ebola virus disease, and other conditions.[10-12,24-26] For HIV, a recent study applied machine learning to EHR data from an academic medical center to predict incident HIV with acceptable precision.[27] Our study extends this work by demonstrating the value of machine learning to identify individuals who merit clinical evaluation for PrEP prospectively, externally, and relative to clinicians' independent PrEP prescribing decisions.

Our study has limitations. First, the HIV prediction model we developed may not be generalizable to organizations that lack the EHR covariates used in our model. However, we used clinical variables commonly embedded in EHRs, including diagnosis codes and common laboratory tests. Notably, ESP, the platform we used for data extraction, is open-source and can interface with any EHR system, so dissemination of HIV prediction models using this platform could facilitate generalizability. Second, our strict definition of incident HIV was intended to exclude all non-incident cases of HIV, but we may have inadvertently excluded some recent HIV infections, which could affect model performance. Third, our model likely underestimated risk in patients with a paucity of EHR data. Fourth, the absence of comprehensive behavioral data in structured EHRs (e.g. data on sexual contacts with HIV-infected partners) precludes comparing our model to CDC indications for PrEP. However, CDC indications for PrEP may have low sensitivity for identifying individuals who acquire HIV.[28] It is possible that some patients identified as potential PrEP candidates by our model may still benefit from using PrEP even if they do not meet CDC criteria. Fifth, race was not recorded for many Atrius Health patients. Our final model included predictor variables for race despite these missing data, suggesting that this model can be applied under real-world conditions where identification of patients' racial characteristics may be incomplete. Sixth, our datasets included few women with incident HIV infection. Developing prediction models using data from populations with more women who acquire HIV might improve model performance for those populations. Seventh, our case-detection algorithms could have misclassified repeated HIV postexposure prophylaxis (PEP) prescriptions as PrEP. We expect that misclassification was uncommon given predominant use of 3-drug PEP at Atrius Health. Eighth, our model had generally lower positive predictive values than HIV risk prediction tools that incorporate patient-reported behavioral data.[29] However, our model had substantially better discrimination, suggesting the usefulness of our model as a first-stage screening tool to prompt clinical evaluations. Finally, our study data was set exclusively in primary care settings. As clinical decision support for emergency department providers can improve PrEP uptake,[30] and because many high-risk patients receive healthcare outside of primary care, such as STI clinics, further development of prediction models for additional healthcare settings is warranted.

In conclusion, our study demonstrates that automated HIV prediction algorithms can harness the data in EHRs to efficiently identify potential PrEP candidates. Additional studies are needed to further optimize these models, integrate them into EHRs at the point-of-care, and evaluate their impact on PrEP prescribing and HIV prevention.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

We would like to thank Victoria Powell, MPH for helping in the preparation of this manuscript and Julia Marcus, PhD for reviewing and providing feedback on our manuscript drafts.

FUNDING SOURCE

This work was supported by: 1) the Harvard University Center for AIDS Research [P30 AI060354], the Providence/Boston Center for AIDS Research [P30AI042853], and the Rhode Island IDeA-CTR [U54GM11567], programs funded by the National Institutes of Health; 2) the National Institute of Mental Health [K23 MH098795]; and 3) the US Centers for Disease Control and Prevention through the STD Surveillance Network [SSuN, CDC-RFA-PS13-1306].

REFERENCES

1. Grant RM, Lama JR, Anderson PL, et al. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. N Engl J Med 2010; 363(27): 2587–99. [PubMed: 21091279]
2. Baeten JM, Donnell D, Ndase P, et al. Antiretroviral prophylaxis for HIV prevention in heterosexual men and women. N Engl J Med 2012; 367(5): 399–410. [PubMed: 22784037]
3. Thigpen MC, Kebaabetswe PM, Paxton LA, et al. Antiretroviral preexposure prophylaxis for heterosexual HIV transmission in Botswana. N Engl J Med 2012; 367(5): 423–34. [PubMed: 22784038]
4. Choopanya K, Martin M, Suntharasamai P, et al. Antiretroviral prophylaxis for HIV infection in injecting drug users in Bangkok, Thailand (the Bangkok Tenofovir Study): a randomised, double-blind, placebo-controlled phase 3 trial. Lancet 2013; 381(9883): 2083–90. [PubMed: 23769234]
5. US Public Health Service. Preexposure Prophylaxis for the Prevention of HIV Infection in the United States - 2014. A Clinical Practice Guideline. Accessed at: https://www.cdc.gov/hiv/pdf/prepguidelines2014.pdf on January 18, 2019.
6. Smith DK, Van Handel M, Grey J. Estimates of adults with indications for HIV pre-exposure prophylaxis by jurisdiction, transmission risk group, and race/ethnicity, United States, 2015. Ann Epidemiol 2018; 12;28(12):850–57 2018. [PubMed: 29941379]

7. Emory University Rollins School of Public Health. AIDSVu. 2019 Accessed at: https://aidsvu.org/resources/#/ on January 18, 2019.

8. Wu H, Mendoza MC, Huang Y-1A, Hayes T, Smith DK, Hoover KW. Uptake of HIV preexposure prophylaxis among commercially insured persons—United States, 2010–2014. Clin Inf Dis 2016; 64(2): 144–49.

9. Krakower DS, Ware NC, Maloney KM, Wilson IB, Wong JB, Mayer KH. Differing Experiences with Pre-Exposure Prophylaxis in Boston Among Lesbian, Gay, Bisexual, and Transgender Specialists and Generalists in Primary Care: Implications for Scale-Up. AIDS Patient Care and STDS 2017; 31(7): 297–304. [PubMed: 28574774]

10. 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.

11. Wallert J, Tomasoni M, Madison G, Held C. Predicting two-year survival versus non-survival after first myocardial infarction using machine learning and Swedish national register data. BMC Med Inform Decis Mak 2017; 17(1): 99. [PubMed: 28679442]

12. Taylor RA, Pare JR, Venkatesh AK, et al. Prediction of In-hospital Mortality in Emergency Department Patients With Sepsis: A Local Big Data-Driven, Machine Learning Approach. Acad Emerg Med 2016; 23(3): 269–78. [PubMed: 26679719]

13. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PLoS One 2017; 12(4): e0174708. [PubMed: 28384212]

14. Sengupta N, Tapper EB. Derivation and Internal Validation of a Clinical Prediction Tool for 30-Day Mortality in Lower Gastrointestinal Bleeding. Am J Med 2017; 130(5): 601 e1–e8.

15. 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.

16. Klompas M, McVetta J, Lazarus R, et al. Integrating clinical practice and public health surveillance using electronic medical record systems. Am J Public Health 2012; 102 Suppl 3: S325–32. [PubMed: 22690967]

17. Bernstein KT, Liu KL, Begier EM, Koblin B, Karpati A, Murrill C. Same-sex attraction disclosure to health care providers among New York City men who have sex with men: implications for HIV testing approaches. Arch Intern Med 2008; 168(13): 1458–64. [PubMed: 18625927]

18. Rucker AJ, Murray A, Gaul Z, Sutton MY, Wilson PA. The role of patient-provider sexual health communication in understanding the uptake of HIV prevention services among Black men who have sex with men. Cult Health Sex 2017;20(7): 761–71. [PubMed: 28929864]

19. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. JAMA 2017; 318(6): 517–8. [PubMed: 28727867]

20. Huang Y-1A, Zhu W, Smith DK, Harris N, Hoover KW. HIV preexposure prophylaxis, by race and ethnicity—United States, 2014–2016.Morb Mortal Wkly Rep 2018; 67(41): 1147–50.

21. Calabrese SK, Earnshaw VA, Krakower DS, et al. A Closer Look at Racism and Heterosexism in Medical Students’ Clinical Decision-Making Related to HIV Pre-Exposure Prophylaxis (PrEP): Implications for PrEP Education. AIDS Behav 2018; 22(4): 1122–38. [PubMed: 29151200]

22. Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. N Engl J Med 2018; 378(11): 981–83. [PubMed: 29539284]

23. Calabrese SK, Magnus M, Mayer KH, et al. "Support Your Client at the Space That They're in": HIV Pre-Exposure Prophylaxis (PrEP) Prescribes' Perspectives on PrEP-Related Risk Compensation. AIDS Patient Care and STDS 2017; 31(4): 196–204. [PubMed: 28414261]

24. Mortazavi BJ, Downing NS, Bucholz EM, et al. Analysis of Machine Learning Techniques for Heart Failure Readmissions. Circ Cardiovasc Qual Outcomes 2016; 9(6): 629–40. [PubMed: 28263938]

25. Weiss JC, Page D, Peissig PL, Natarajan S, McCarty C. Statistical Relational Learning to Predict Primary Myocardial Infarction from Electronic Health Records. Proceedings of the Innovative Applications of Artificial Intelligence Conference / sponsored by the American Association for

Artificial Intelligence Innovative Applications of Artificial Intelligence Conference 2012; 2012: 2341–47.

26. Wiens J, Shenoy ES. Machine Learning for Healthcare: On the Verge of a Major Shift in Healthcare Epidemiology. Clin Infect Dis 2018; 66(1): 149–53. [PubMed: 29020316]
27. 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): 160–66. [PubMed: 29084046]
28. Hoots BE, Finlayson T, Nerlander L, Paz-Bailey G. Willingness to Take, Use of, and Indications for Pre-exposure Prophylaxis Among Men Who Have Sex With Men-20 US Cities, 2014. Clin Infect Dis 2016; 63(5): 672–77. [PubMed: 27282710]
29. 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]
30. 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 and STDS 2018; 32(5): 202–7. [PubMed: 29672136]

RESEARCH IN CONTEXT

Evidence before this study

We searched PubMed using combinations of the search terms "HIV", "preexposure prophylaxis", "preexposure prophylaxis", "PrEP", "risk prediction", "risk score", "clinical prediction rule", "prediction model", "risk assessment tool", "predictive analytics", and "machine learning" for all articles published on or before January 9, 2019 (the date of our final search). Studies have demonstrated that HIV preexposure prophylaxis (PrEP) can decrease HIV incidence in priority populations. However, PrEP uptake in the United States has been limited thus far. One of the barriers to scale-up PrEP is that few primary care providers have prescribed PrEP, in part because these providers do not routinely ask their patients about sexual and substance use behaviors to determine their eligibility for PrEP use. The United States Preventive Services Task Force has identified a need to develop and validate tools to identify persons at increased risk for HIV acquisition as a way to improve PrEP provision. Our literature search revealed that there are few validated tools to help providers with HIV risk assessments. Moreover, existing risk assessment tools have suboptimal predictive performance and require providers to manually calculate risk scores for individual patients, both of which limit their utility in clinical practice.

Added value of this study

We used machine learning algorithms to develop and validate an automated model to predict incident HIV infections within an ambulatory group practice in Massachusetts using electronic health record data. The model was validated prospectively at this practice and externally using data from a community health center specializing in sexual healthcare in Boston. Cohorts included 1,155,966 ambulatory practice patients from 2007-2015 for model development (including 150 [<0·1%] patients with incident HIV), 537,257 patients in 2016 for prospective validation (16 [>0·1%] with incident HIV), and 33,404 community health center patients from 2011-2016 (423 [1·3%] with incident HIV). The prediction model had a cross-validated area under the curve (cv-AUC) of 0·86 for identifying incident HIV infections in the development cohort, 0·91 on prospective validation, and 0·77 on external validation. The model could also identify patients independently prescribed PrEP by clinicians at the ambulatory practice (cv-AUC 0·94) or the community health center (cv-AUC 0·79). Patients' HIV risk scores at the ambulatory practice increased steeply at the 98th percentile of scores. We defined patients with scores above this threshold as potential PrEP candidates and prospectively identified 9,515/536,384 (1·8%) new PrEP candidates at the ambulatory practice in 2016. If clinicians had discussed PrEP with all patients with risk scores above our cut-off and prescribed PrEP to those who indicated current high-risk behaviors, our model could have helped avert nearly 40% of the new HIV infections at Atrius Health in 2016.

Implications of all the available evidence

Automated algorithms can generate models that efficiently identify persons at increased risk for HIV acquisition based on their electronic health records profiles. Model performance for identifying high-risk individuals in distinct populations could potentially be further improved by model building directly upon data from specific target populations. Integrating HIV risk prediction models into primary care and alerting providers about patients who merit clinical evaluations for PrEP use could improve prescribing and prevent new HIV infections.

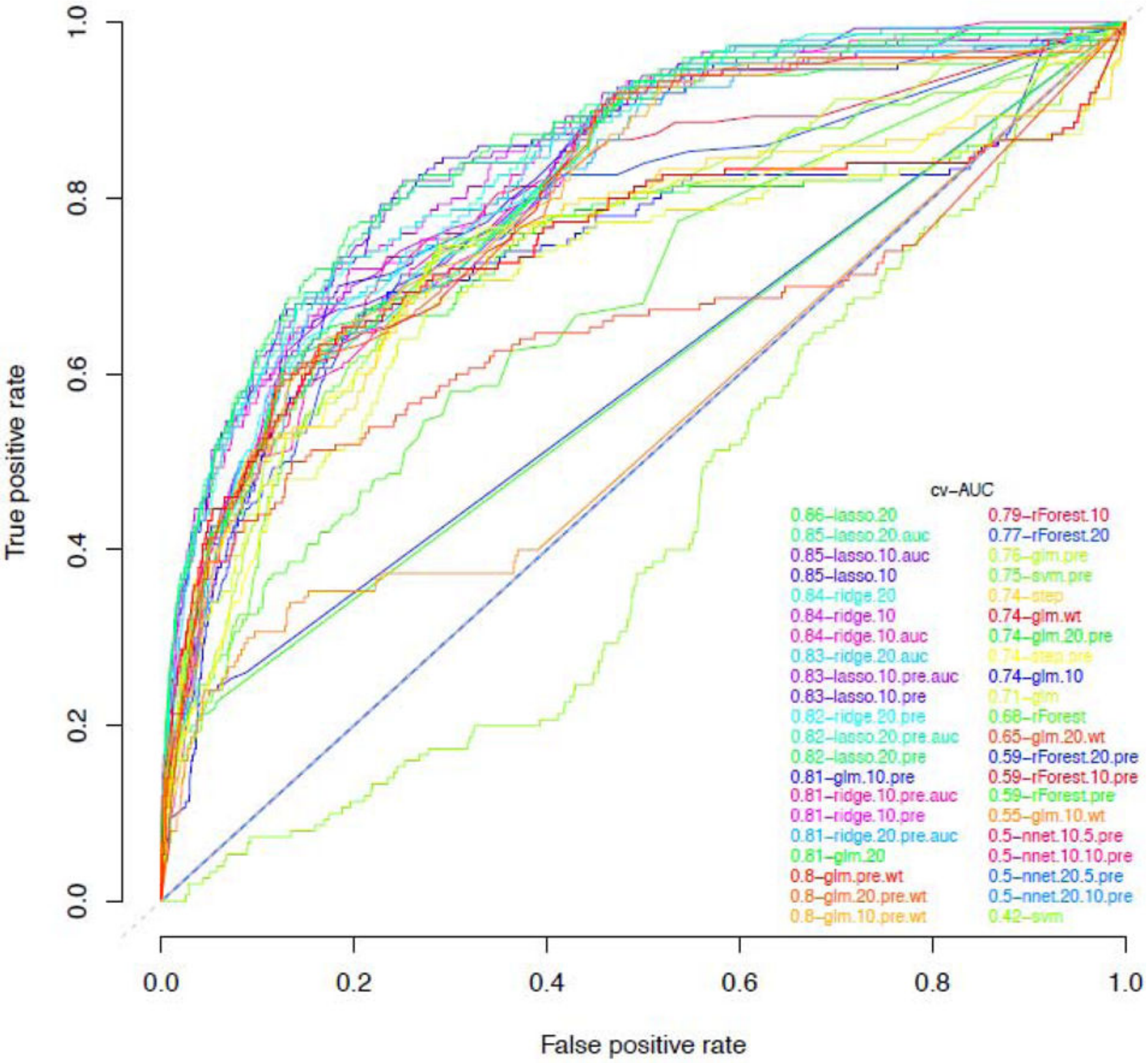

Figure 1: Weighted cross-validated area under the receiver-operating curve (cv-AUC) for 42 candidate prediction algorithms fit on cases (n=150) and controls (n=7,466) in the development cohort - Atrius Health, 2007-2015.

LASSO, least absolute shrinkage and selection operator; ridge, ridge regression; nnet, neural networks; glm, generalized linear model (logistic regression); step, logistic regression with stepwise backwards selection; rForest, random forest; svm, support vector machines. Algorithm abbreviations with “pre” denotes use of preselected covariates, with “auc” denotes AUC loss function instead of deviance loss, with “wt” denotes weighted regression, with “10” denotes undersampling with approximately 1:10 ratio of cases to controls, and with “20” denotes undersampling with approximately 1:20 ratio of cases to controls. For neural nets, the first number denotes the ratio of cases to controls, and the second number denotes the number of nodes in the network’s single hidden layer, e.g., 20·5 indicates a 1:20 case control ratio, with 5 nodes in the network’s hidden layer (appendix p 6).

Table 1:

Demographic characteristics, incident HIV infections, and PrEP use in the development, prospective validation, and external validation cohorts.

	Development Cohort Atrius Health, 2007-2015		Prospective Validation Cohort Atrius Health, 2016	External Validation Cohort Fenway Health, 2011-2016
Parameter	**Total cohort (n=1,155,966)**	**Controls[a] (n=7,466)**	**Total cohort (n=537,257)**	**Total cohort (n=33,404)**
Age (years), mean (SD)[b]	35·0 (22·2)	44·7 (18)	39·1 (23·3)	34·5 (12·3)
Gender, n (%)				
Male	495,871 (42·9)	5967 (79·9)	228,239 (42·5)	20,796 (62·3)
Female	658,351 (57·0)	1499 (20·1)	309,010 (57·5)	10,371 (31·0)
Transgender or gender non-conforming[c]	--	--	--	2,237 (6·7)
Unknown	1744 (0·2)	--	8 (< 0·1)	--
Race/ethnicity, n (%)				
White	694,124 (60·1)	5555 (74·4)	390,353 (72·7)	22,826 (68·3)
Black	60,239 (5·2)	655 (8·8)	37,147 (6·9)	2,706 (8·1)
American Indian or Alaska Native	1,048 (0·1)	8 (0·1)	539 (0·1)	74 (0·2)
Asian	66,810 (5·8)	352 (4·7)	34,192 (6·4)	2,388 (7·1)
Native Hawaiian and Other Pacific Islander	427 (< 0·1)	1 (< 0·1)	195 (< 0·1)	125 (0·4)
Other	38,645 (3·3)	660 (8·8)	21,248 (4·0)	3,409 (10·2)
Hispanic or Latino	33,636 (2·9)	235 (3·1)	17,426 (3·2)	1,876 (5·6)
Unknown	261,037 (22·6)	--	36,157 (6·7)	--
At least 1 EHR predictor variable suggestive of HIV risk[d], n (%)	399,385 (34·5)	n/a	245,459 (45·7)	n/a
Incident HIV, n (%)	150 (< 0·1)	n/a	16 (< 0·1)	423 (1·3)
PrEP use, n (%)	90 (< 0·1)	n/a	128 (< 0·1)	1,813 (5·4)

EHR, electronic health records; PrEP, preexposure prophylaxis; n/a, not applicable.

[a] To create a set of control patients for algorithm development, 0·7% of HIV-uninfected males per year and 0·11% of HIV-uninfected female controls per year were sampled from among the 399,385 patients with EHR data suggestive of HIV risk. These proportions were chosen to yield approximately 50 controls per case.

[b] Age as of the beginning of the study period or, for those patients who had not yet established care as of this date, as of the date of their first documented EHR data element.

[c] Data not available at Atrius Health.

[d] Data on patients with at least 1 EHR predictor variable suggestive of HIV infection are not shown for Fenway Health, as the entire patient population at Fenway Health was used for validation studies given the high HIV incidence in this population.

Table 2:

Prevalence of predictor variables for patients with incident HIV infection versus controls in the development cohort, and LASSO model coefficients for each predictor variable.

Electronic health record predictor variables[a]	Incident HIV (n=150)	Controls (n=7,466)	Coefficient[b]
Diagnosis codes, n (%)			
Syphilis of any site or stage except late latent	6 (4·0)	5 (0·1)	1·00
HIV counseling in prior 2 years	8 (5·3)	26 (0·3)	1·10
Contact with or exposure to venereal disease	15 (10·0)	139 (1·9)	0·29
Laboratory tests			
Number of positive gonorrhea tests in prior 2 years, mean (SD)	0·04 (0·23)	0·00 (0·02)	3·07
Number of Chlamydia tests, mean (SD)	0·00 (0·0)	0·00 (0·03)	−0·15
Number of HIV tests, mean (SD)	0·81 (1·71)	0·18 (0·62)	0·12
Number of HIV ELISA tests, mean (SD)	0·61 (1·35)	0·15 (0·54)	0·16
Number of HIV tests in prior 2 years, mean (SD)	0·44 (0·97)	0·09 (0·34)	0·23
Number of HIV RNA tests in prior year, mean (SD)	0·05 (0·40)	0·00 (0·02)	0·15
Testing for acute HIV[c], n (%)	7 (4·7)	7 (0·1)	1·82
Testing for acute HIV[c] in prior 2 years, n (%)	4 (2·7)	2 (< 0·1)	0·16
Prescriptions, n (%)			
Intramuscular penicillin G benzathine	8 (5·3)	2 (< 0·1)	1·80
Intramuscular penicillin G benzathine in prior year	5 (3·3)	0 (0·0)	1·36
Intramuscular penicillin G benzathine in prior 2 years	5 (3·3)	1 (< 0·1)	0·21
Buprenorphine and naloxone in prior 2 years	2 (1·3)	26 (0·3)	0·20
Registration data			
Years of prior electronic health records data, mean (SD)	2·74 (2·72)	3·92 (2·68)	−0·07
At least 1 year of prior electronic health records data, n (%)	92 (61·3)	6153 (82·4)	−0·63
At least 2 years of prior electronic health records data, n (%)	72 (48·0)	5230 (70·1)	−0·40
Any data on primary language, n (%)	129 (86·0)	7145 (95·7)	−0·08
English as primary language, n (%)	114 (76·0)	6778 (90·8)	−0·42
Black race, n (%)	51 (34·0)	655 (8·8)	1·06
White race, n (%)	55 (36·7)	5555 (74·4)	−0·66
Male gender, n (%)	120 (80)	5967 (79·9)	1·87

ELISA, enzyme-linked immunosorbent assay; RNA, ribonucleic acid.

[a] The variables shown are those included in the final LASSO algorithm.

[b] To calculate an HIV risk prediction score, the value of each variable is multiplied by its coefficient and the products are then summed to generate the risk score on the logit scale. Binary variables are assigned a value of 1 if affirmative and 0 if non-affirmative.

[c] Testing for acute HIV defined as HIV RNA testing among individuals without evidence of HIV infection.

Table 3:

Number of patients with HIV risk scores above specific percentile thresholds in the general population[a] and among those with incident HIV infection or PrEP use in the three study cohorts.

		Atrius Health, 2007-2015 (n=1,154,964)			Atrius Health, 2016 (n=536,528)			Fenway Health, 2011-2016 (n=30,938)		
Percentile of HIV risk score[b]	**Risk score (out of 100,000)**	**General population (n=1,154,724) n (%)**	**Incident HIV infection (n=150) n (%)**	**PrEP use (n=90) n (%)**	**General population (n=536,384) n (%)**	**Incident HIV infection (n=16) n (%)**	**PrEP use (n=128) n (%)**	**General population (n=28,702) n (%)**	**Incident HIV infection (n=423) n (%)**	**PrEP use (n=1,813) n (%)**
10%	0	399,385 (34·6)	150 (100·0)	90 (100)	245,459 (45·8)	16 (100)	128 (100·0)	28,702 (100·0)	423 (100·0)	1,813 (100·0)
20%	0	399,385 (34·6)	150 (100·0)	90 (100)	245,459 (45·8)	16 (100)	128 (100·0)	28,702 (100·0)	423 (100·0)	1,813 (100·0)
30%	0	399,385 (34·6)	150 (100·0)	90 (100)	245,459 (45·8)	16 (100)	128 (100·0)	28,702 (100·0)	423 (100·0)	1,813 (100·0)
40%	0	399,385 (34·6)	150 (100·0)	90 (100)	245,459 (45·8)	16 (100)	128 (100·0)	28,702 (100·0)	423 (100·0)	1,813 (100·0)
50%	0	399,385 (34·6)	150 (100·0)	90 (100)	245,459 (45·8)	16 (100)	128 (100·0)	28,702 (100·0)	423 (100·0)	1,813 (100·0)
60%	0	399,385 (34·6)	150 (100·0)	90 (100)	245,459 (45·8)	16 (100)	128 (100·0)	28,702 (100·0)	423 (100·0)	1,813 (100·0)
70%	1	342,911 (29·7)	144 (96·0)	90 (100)	173,639 (32·4)	16 (100)	128 (100·0)	28,125 (98·0)	423 (100·0)	1,810 (99–8)
80%	2	225,988 (19·6)	142 (94·7)	90 (100)	129,767 (24·2)	16 (100)	128 (100·0)	21,015 (73·2)	415 (98·1)	1,792 (98·8)
90%	8	115,269 (10·0)	116 (77·3)	88 (98)	48,533 (9·0)	15 (94)	115 (89·8)	16,023 (55·8)	386 (91·3)	1,721 (94·9)
91%	9	103,623 (9·0)	107 (71·3)	88 (98)	47,708 (8·9)	15 (94)	114 (89·0)	15,690 (54·7)	385 (91·0)	1,706 (94·1)
92%	10	92,278 (8·0)	105 (70·0)	88 (98)	40,961 (7·6)	13 (81)	108 (84·4)	14,861 (51·8)	370 (87·5)	1,672 (92·2)
93%	11	76,263 (6·6)	105 (70·0)	88 (98)	34,041 (6·3)	13(81)	106 (82·8)	14,016 (48·8)	362 (85·6)	1,623 (89·5)
94%	12	67,508 (5·8)	103 (68·7)	88 (98)	28,572 (5·3)	12 (75)	102 (79·7)	13,045 (45·4)	351 (83·0)	1,569 (86·5)
95%	13	57,470 (5·0)	101 (67·3)	88 (98)	24,413 (4·6)	10 (63)	99 (77·3)	11,732 (40·9)	340 (80·4)	1,533 (84·6)
96%	15	43,312 (3·8)	97 (64·7)	85 (94)	18,854 (3·5)	8 (50)	90 (70·3)	6,961 (24·3)	225 (53·2)	1,125 (62·1)
97%	18	34,247 (3·0)	85 (56·7)	83 (92)	15,101 (2·8)	8 (50)	82 (64·0)	5,997 (20·9)	207 (48·9)	1,039 (57·3)
98%	25	23,018 (2·0)	76 (50·7)	71 (79)	9,515 (1·8)	6 (38)	62 (48·4)	4,385 (15·3)	196 (46·3)	851 (46·9)

		Atrius Health, 2007-2015 (n=1,154,964)			Atrius Health, 2016 (n=536,528)			Fenway Health, 2011-2016 (n=30,938)		
Percentile of HIV risk score[b]	**Risk score (out of 100,000)**	**General population (n=1,154,724) n (%)**	**Incident HIV infection (n=150) n (%)**	**PrEP use (n=90) n (%)**	**General population (n=536,384) n (%)**	**Incident HIV infection (n=16) n (%)**	**PrEP use (n=128) n (%)**	**General population (n=28,702) n (%)**	**Incident HIV infection (n=423) n (%)**	**PrEP use (n=1,813) n (%)**
99%	32	11,352 (1·0)	62 (41·3)	56 (62)	5,192 (1·0)	4 (25)	53 (41·4)	2,284 (8·0)	131 (31·0)	672 (37·1)

PrEP, preexposure prophylaxis.

[a] Excludes patients already diagnosed with HIV infection for all study cohorts, for whom HIV risk scores would not be meaningful.

[b] HIV risk scores were calculated by applying the LASSO model to the development sample.

Table 4:

Performance of LASSO algorithm for detecting incident HIV infection in the development cohort.

Percentile of HIV risk score used to define test positivity	Risk score (out of 100,000)	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)[a]
10%	0	100	65·4	0·04	100
20%	0	100	65·4	0·04	100
30%	0	100	65·4	0·04	100
40%	0	100	65·4	0·04	100
50%	0	100	65·4	0·04	100
60%	0	100	65·4	0·04	100
70%	1	96·0	70·3	0·04	100
80%	2	94·7	80·4	0·06	100
90%	8	77·3	90·0	0·10	100
91%	9	71·3	91·0	0·10	100
92%	10	70·0	92·0	0·11	100
93%	11	70·0	93·1	0·13	100
94%	12	68·7	94·2	0·15	100
95%	13	67·3	95·0	0·18	100
96%	15	64·7	96·3	0·22	100
97%	18	56·7	97·0	0·25	100
98%	25	50·7	98·0	0·33	100
99%	32	41·3	99·0	0·54	100

[a] High negative predictive values reflect the low HIV incidence in the population.