

¹Division of Infectious Diseases, Beth Israel Deaconess Medical Center, Boston, MA; ²The Fenway Institute, Fenway Health, Boston, MA; ³Harvard Medical School, Boston, MA; ⁴Department of Population Medicine, Harvard Medical School, Boston, MA; ⁵Massachusetts Department of Public Health, Boston, MA; ⁶Department of Pediatrics, Boston Medical Center, Boston, MA; ⁷Atrius Health, Boston, MA; ⁸Department of Health Services, Policy and Practice, Brown University, Providence, RI; ⁹Division of Infectious Diseases, Brigham and Women's Hospital, Boston, MA

Background

- To maximize the population-level impact of pre-exposure prophylaxis (PrEP), healthcare organizations need tools to identify persons at risk for HIV infection
- We hypothesized that electronic health record (EHR) data could be used to identify patients at increased risk for acquiring HIV, who might thus be candidates for PrEP

Methods

- We developed and evaluated automated algorithms to predict incident HIV using EHR data from Fenway Health, a community health center in Boston specializing in health care for sexual and gender minorities
- EHR data was extracted for 168 variables possibly associated with incident HIV for all patients with ≥ 1 clinical encounter during 2011-2016
- Variables included demographics, laboratory tests/results, diagnosis codes, and prescriptions
- Algorithms developed using machine learning methods (LASSO, Ridge Regression, Random Forest) and Generalized Linear Models were used to estimate risk for all patients
- Algorithms were trained using 2011-2015 data and validated using 2016 data
- Algorithm performance was assessed using 10-fold, cross-validated area under the receive operator curves (cv-AUC), sensitivity, specificity, and positive predictive value (PPV), and negative predictive value (NPV)

Results

- 423 of 33,404 patients had incident HIV, including 71 of 18,275 patients in 2016 (Table 1)
- LASSO was the best-performing algorithm with a cv-AUC of 0.82 (Figure 1)
- A step increase in risk scores was detected above the 98th percentile of scores in the general population; these patients represent potential PrEP candidates (Figure 2)
- Using the top 20% of patient risk scores, LASSO had a sensitivity of 73%, specificity of 81%, PPV of 1.5% for predicting incident HIV in 2016 (Figure 3)

Table 1. Demographics of Fenway Health cohort, 2011-2016

Parameter	Total Cohort (n=33,404)
Age, mean (SD), y	34.5 (12.3)
Gender, No. (%)	
Male	20,796 (60.4)
Female	10,371 (30.1)
Transgender or gender non-conforming	2,237 (6.5)
Race/ethnicity, No. (%)	
White	22,826 (65)
Black	2,706 (8)
American Indian or Alaska Native	74 (0.2)
Asian	2,388 (7)
Native Hawaiian and Other Pacific Islander	125 (0.4)
Other	3,409 (10)
Hispanic or Latino	1,876 (5)
Incident HIV	423 (1.2)
PrEP Use	1,813 (5.3)

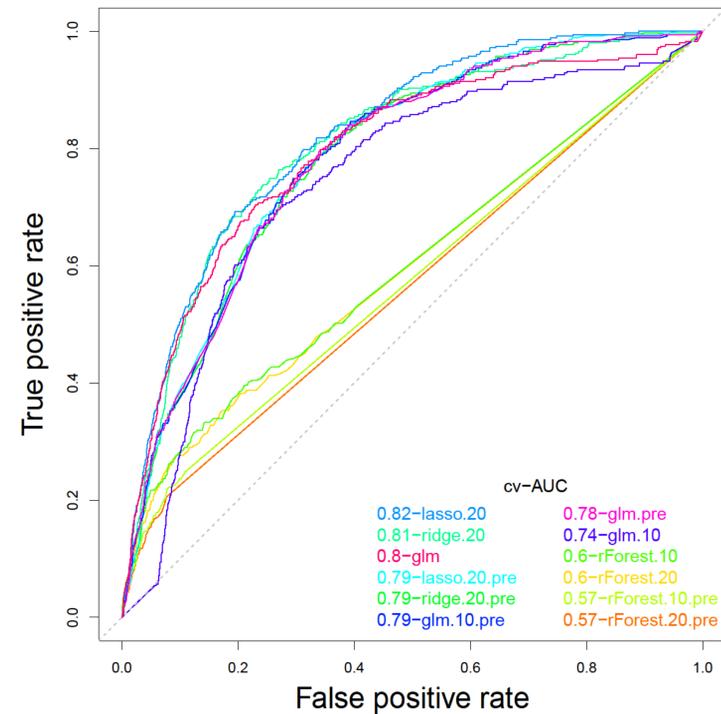


Figure 1. Weighted cross-validated area under the receiver-operating curve (cv-AUC) for candidate prediction algorithms fit on cases (n=351) and controls (73,644 person-years contributed by n=28,790 unique patients during 2011-2015). Ridge, ridge regression; rForest, Random Forest; glm, Generalized Linear Models. Legend suffixes: "20" and "10" indicate controls were under sampled to produce approximately a 1:20 or 1:10 class ratio; "pre" indicates algorithms were fit on a pre-selected subset of covariates instead of all available covariates.

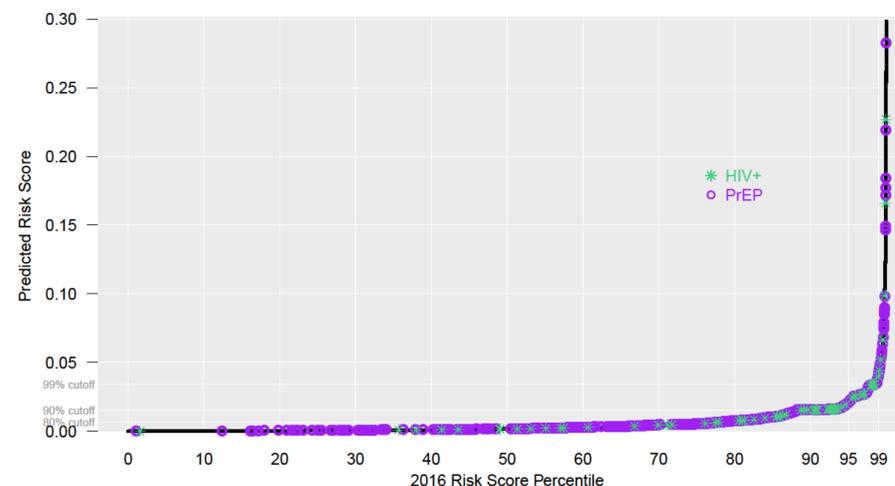


Figure 2. Distributions of HIV risk prediction scores for patients in the general population (black line) and those with incident HIV or PrEP use. A steep increase in risk scores occurs at the 98th percentile of risk scores.

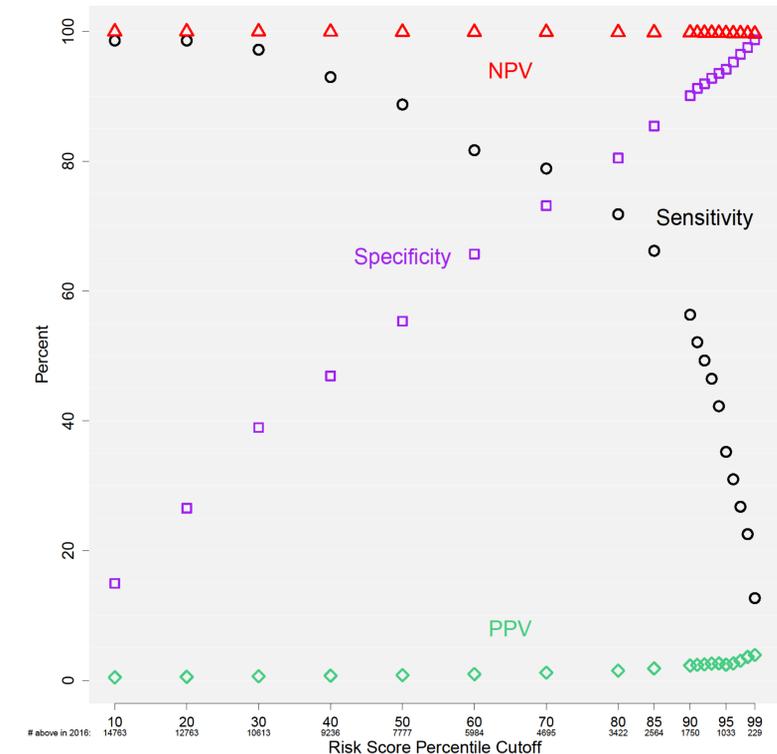


Figure 3. Sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) of the best-performing algorithm (LASSO) at identifying patients with incident HIV infection in 2016 at varying HIV risk score cut-offs. The Figure illustrates trade-offs in sensitivity and specificity at different cut-off values. PPV was generally low but increased at the highest range of HIV risk scores.

Conclusions

- Automated algorithms that integrate EHR data have favorable properties as population-level screening tools to identify patients at high risk for HIV infection who are not already using PrEP
- PPV for incident HIV was generally low. However, at high HIV risk scores, the PPV's were comparable to those of currently available manual HIV risk prediction tools (e.g. ARCH-MSM [Smith JAIDS 2012]).
- These algorithms offer an efficient means of reducing missed opportunities to provide PrEP to patients most likely to acquire HIV
- Future studies to integrate these algorithms into routine care and evaluate their impact on PrEP use and HIV incidence are warranted.

Acknowledgements

- Victoria Powell, MPH for assistance with presentation of data. 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). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the US Centers for Disease Control and Prevention.