Appendix material for "Development and validation of an automated HIV prediction algorithm to identify candidates for preexposure prophylaxis"

Appendix Table 1. Electronic health record variables used for development and validation of HIV prediction model.

Predictor variables		
Demographics		
Age		
Sex or gender[a]		
Race/ethnicity[a,b]		
Primary language[a,b]		
Number of years of electronic health records data recorded[b]		
Clinical encounters		
Number of clinical encounters		
Number of clinical encounters in the prior year		
Number of clinical encounters in the prior 2 years		
Prescriptions		
Ceftriaxone injection 125 mg or 250 mg intravenous or intramuscular by year		
Methadone by year		
Buprenorphine plus naloxone prescription by year[a,b]		
Sildenafil, tadalafil, or vardenafil prescription by year		
Penicillin G benzathine 2.4 million units injection by year[b]		
Azithromycin 1 g orally by year		
Tenofovir disoproxil fumarate/emtricitabine prescription for use as PrEP, defined as ≥2 prescriptions written ≥2 months apart during years without meeting case-detection criteria for HIV-infection[c] or chronic hepatitis B infection[d], by year[e]		
Diagnoses	**ICD-9 codes**	**ICD-10 codes**
Abnormal anal cytology, anal dysplasia, or anal carcinoma in situ	796·7x, 569·44, 230·5, 230·6	D01·3
Alcohol dependence	305·00, 305·01, 305·02, 303·90, 303·91, 303·92	F10·2
Amphetamine dependence	304·4x	F15·x
Anal syphilis	091·1	A51·1
Anogenital condyloma	078·11	A63·0
Anorectal ulcer	569·41	K62·6
Anorexia nervosa	307·1	F50·0
Bulimia nervosa[a]	307·51	F50·2
Chancroid	099·0	A57
Chlamydial infection of anus and rectum	099·52	A56·3
Chlamydial infection of pharynx	099·51	A56·4
Cocaine dependence	304·2x	F14·x
Counseling for child sexual abuse	V61·21	Z61·4, Z61·5
Contact with or exposure to venereal disease[a,b]	V01·6	Z20·2, Z20·6
Gender identity disorders[a]	302·0	F64·8, F64·9, F66·0
Eating disorder NOS	307·50	F50·9
Foreign body in anus	937	T18·5
Genital herpes	054·1x	A60·0
Gonococcal infection of anus and rectum	098·7	A54·6
Gonococcal pharyngitis	098·6	A54·5
Granuloma inguinale	099·2	A58
Herpes simplex with complications	054·8, 054·9, 054·79	A60·1, A60·9
High risk sexual behavior[a]	V69·2	Z72·5
HIV infection	042·1	B20·x to B24·x
HIV counseling[a,b]	V65·44	Z71·7
Lymphoma granuloma venereum	099·1	A55
Nongonococcal urethritis[a]	099·4x	N34·1
Opioid dependence	304·0x	F11·x
Pelvic inflammatory disease	614, 614·1-514·3, 614·5, 614·9	N72, N73·0, N73·2, N73·5, N73·9
Sedative, hypnotic, or anxiolytic dependence	304·1x	F13·x
Syphilis of any site or stage except late latent[a,b]	091·x, 092·x, 093·x, 094·x, 095·x, 097·9	A51·x, A52·0, A52·7
Trans-sexualism	302·50, 302·51, 302·52, 302·53	F64·0, F64·1
Unspecified drug dependence	304·9x	F19·x
Unspecified sexually transmitted disease		A63·x, A64

Predictor variables
Laboratory tests and results
Number of Chlamydia tests[b]
Number of Chlamydia tests in the prior year
Number of Chlamydia tests in the prior 2 years[a]
Number of positive Chlamydia tests
Number of positive Chlamydia tests in the prior year
Number of positive Chlamydia tests in the prior 2 years
Number of rectal site Chlamydia tests
Number of rectal site Chlamydia tests in the prior year
Number of rectal site Chlamydia tests in the prior 2 years
Number of rectal site positive Chlamydia tests
Number of rectal site positive Chlamydia tests in the prior year
Number of rectal site positive Chlamydia tests in the prior 2 years
Number of oropharyngeal site Chlamydia tests
Number of oropharyngeal site Chlamydia tests in the prior year
Number of oropharyngeal site Chlamydia tests in the prior 2 years
Number of positive oropharyngeal site Chlamydia tests
Number of positive oropharyngeal site Chlamydia tests in the prior year
Number of positive oropharyngeal site Chlamydia tests in the prior 2 years
Serological testing for lymphogranuloma venereum
Number of Gonorrhea tests
Number of Gonorrhea tests in the prior year
Number of Gonorrhea tests in the prior 2 years[a]
Number of positive gonorrhea tests[a]
Number of positive gonorrhea tests in the prior year
Number of positive gonorrhea tests in the prior 2 years[b]
Number of rectal site gonorrhea tests
Number of rectal site gonorrhea tests in the prior year
Number of rectal site gonorrhea tests in the prior 2 years
Number of rectal site positive gonorrhea tests
Number of rectal site positive gonorrhea tests in the prior year
Number of rectal site positive gonorrhea tests in the prior 2 years
Number of oropharyngeal site gonorrhea tests
Number of oropharyngeal site gonorrhea tests in the prior year
Number of oropharyngeal site gonorrhea tests in the prior 2 years
Number of positive oropharyngeal site gonorrhea tests
Number of positive oropharyngeal site gonorrhea tests in the prior year
Number of positive oropharyngeal site gonorrhea tests in the prior 2 years
Number of Syphilis tests
Number of Syphilis tests in the prior year
Number of Syphilis tests in the prior 2 years
Anal cytology testing
Number of hepatitis C virus antibody tests
Number of hepatitis C virus antibody tests in the prior year[a]
Number of hepatitis C virus antibody tests in the prior 2 years
Number of hepatitis C virus RNA tests
Number of hepatitis C virus RNA tests in the prior year
Number of hepatitis C virus RNA tests in the prior 2 years
Positive hepatitis C virus antibody or RNA test
Year of first positive hepatitis C virus antibody test
Year of first positive hepatitis C virus RNA test
Number of hepatitis B virus surface antigen tests in the prior year
Number of hepatitis B virus surface antigen tests in the prior 2 years
Number of hepatitis B virus surface antigen tests, ever
Number of hepatitis B virus DNA tests in the prior year
Number of hepatitis B virus DNA tests in the prior 2 years
Number of hepatitis B virus DNA tests, ever
Reactive hepatitis B surface antigen or detectable hepatitis B DNA
Year of first positive for hepatitis B virus infection
Number of HIV tests[b]
Number of HIV tests in the prior year
Number of HIV tests in the prior 2 years[a,b]
Number of HIV ELISA tests[b]
Number of HIV ELISA tests in the prior year
Number of HIV ELISA tests in the prior 2 years
Number of HIV Western Blot tests
Number of HIV Western Blot tests in the prior year
Number of HIV Western Blot tests in the prior 2 years
Number of HIV RNA tests[a]

Predictor variables
Number of HIV RNA tests in the prior year[b]
Number of HIV RNA tests in the prior 2 years
Testing for acute HIV, defined as HIV RNA testing and not meeting case-detection criteria for HIV infection[b,c]
Testing for acute HIV, defined as HIV RNA testing and not meeting case-detection criteria for HIV infection[c], in the prior 2 years[b]
Combinations of Variables
Positive case-detection criteria for HIV infection[c]
First year of meeting positive case-detection criteria for HIV infection[c]
Positive case-detection criteria for acute hepatitis B infection[d]
First year of meeting positive case-detection criteria for acute hepatitis B infection[d]
Positive case-detection criteria for acute hepatitis C infection[f]
Positive case-detection criteria for Syphilis[g]
Positive case-detection criteria for Syphilis[g] in the prior year
Positive case-detection criteria for Syphilis[g] in the prior 2 years
Female sex or gender with any variables indicating testing, diagnosis, or treatment for gonorrhea, chlamydia or syphilis, or with diagnosis of unspecified sexually transmitted disease or pelvic inflammatory disease[a]
Trans-sexualism or gender identity disorders
Opioid dependence or Buprenorphine plus naloxone prescription or methadone prescription or amphetamine dependence or cocaine dependence

NOS, not otherwise specified; ICD-9 and ICD-10, International Classification of Diseases, Ninth and Tenth Revisions, respectively; RNA, ribonucleic acid; ELISA, enzyme-linked immunosorbent assay.

[a]Predictors of incident HIV infection pre-selected by clinical experts.

[b]Predictors of incident HIV infection included in the final LASSO model.

[c]HIV infection defined as any of the following: 1) a positive Western Blot, Multispot, or Geenius test result; 2) a positive HIV Antigen/Antibody test AND a positive HIV ELISA; 3) HIV RNA Viral Load over 200 copies/mL; 4) HIV Qualitative PCR; 5) 2 or more ICD codes for HIV and a history of prescription for 3 or more HIV medications, ever; 6) HIV on the medical problem list and a history of prescription for 3 or more HIV medications, ever; 7) concurrent prescriptions for 3 or more different antiretroviral medications for at least 1 month.

[d]Acute hepatitis B infection defined as any of the following: 1) One of the following: a) diagnosis code for jaundice (not of newborn), b) Alanine aminotransferase (ALT) > than 5 times the upper limit of normal, or c)aspartate aminotransferase (AST) > than 5 times the upper limit of normal, AND IgM antibody to Hepatitis B core Antigen "reactive" within a 14 day period; 2) One of the following: a) diagnosis code for jaundice (not of newborn), b) ALT > than 5 times the upper limit of normal, or c) AST > than 5 times the upper limit of normal, AND one of the following: a) total bilirubin > 1·5 or b) calculated bilirubin > 1·5, AND Hepatitis B Surface Antigen within a 21-day period, AND no prior positive result for hepatitis B Surface Antigen or hepatitis B viral DNA, AND no codes for diagnosis for chronic hepatitis B in the past; 3) One of the following: a) diagnosis code for jaundice (not of newborn), b) ALT > than 5 times the upper limit of normal, or c)AST > than 5 times the upper limit of normal, AND one of the following: a) total bilirubin > 1·5 or b) calculated bilirubin > 1·5, AND hepatitis B viral DNA within a 21-day period, AND no prior positive result for hepatitis B Surface Antigen or hepatitis viral DNA, AND no codes for diagnosis of chronic hepatitis B in the past; 4) hepatitis B Surface Antigen "reactive" with record of hepatitis B Surface Antigen "non-reactive" within the prior 12 months, AND no prior positive result for hepatitis B Surface Antigen or hepatitis viral DNA, AND no codes for diagnosis of chronic hepatitis B in the past.

[e]Prescriptions for PrEP were not included as a predictor variable in our model. Data on PrEP use was collected so we could examine model performance at identifying patients who were prescribed PrEP by clinicians.

[f]Acute hepatitis C infection defined as any of the following: 1) any one of the following: a) diagnosis code for jaundice, b) ALT > 200, or c) total bilirubin > 1·5, AND reactive hepatitis C ELISA, AND all of the following if completed: a) hepatitis C signal cutoff ratio ≥3·8, b) positive hepatitis C recombinant immunoblot assay (RIBA), and c) detectable hepatitis C RNA, all the above criteria within a 28-day period, AND no prior positive: a) hepatitis C ELISA, b) hepatitis C RIBA, or c) detectable hepatitis C RNA, AND no prior diagnosis code for chronic hepatitis C; 2) any one of the following: a) diagnosis code for jaundice, b) ALT > 200, or c) total bilirubin > 1·5, AND

detectable hepatitis C RNA and all of the following if done: a) hepatitis C signal cutoff ratio ≥3.8, and b) hepatitis C RIBA, all of the above criteria within a 28-day period, AND no prior positive: a) hepatitis C ELISA, b) hepatitis C RIBA, or c) detectable hepatitis C RNA, AND no prior diagnosis code for chronic hepatitis C; 3) detectable hepatitis C RNA AND in the prior 12 months a recorded negative: a) hepatitis C ELISA, or b) detectable hepatitis C RNA, AND no prior positive: a) hepatitis C ELISA, b) hepatitis C RIBA, or c) detectable hepatitis C RNA, AND no prior diagnosis code for chronic hepatitis; 4) reactive hepatitis C ELISA AND in the prior 12 months a recorded negative: a) hepatitis C ELISA or b) hepatitis C RNA, AND no prior positive: a) hepatitis C ELISA, b) hepatitis C RIBA, or c) detectable hepatitis C RNA, AND no prior diagnosis code for chronic hepatitis.

[g]Active syphilis defined as meeting any of the following three criteria: 1) (Diagnosis code for syphilis or positive Treponema pallidum IgM test) and an order or prescription for at least one of the following antibiotics within a 14-day period: a) penicillin G; b) doxycycline for > 7 day duration, or c) ceftriaxone dosed at ≥ 1 gram; 2) Serum reactive plasma regain (RPR) or Venereal Disease Research Laboratory (VDRL) test value greater than or equal to 1:8 and any of the following: a) Treponema pallidum particle agglutination (TPPA) test with result "reactive" ever in the past and up to 1 month following the positive RPR or VDRL; b). fluorescent treponemal antibody absorption (FTA-ABS) test with result "reactive" ever in the past and up to 1 month following the positive RPR or VDRL; or c) Treponema pallidum IgG test (TP-IGG) test with result “positive” or “reactive” ever in the past and up to 1 month following positive RPR or VDRL; or 3) Positive cerebrospinal fluid test for syphilis, including any of the following: a) VDRL-CSF value “reactive” or ≥ 1:14; b) TPPA-CSF with result “reactive” or “positive” or equivalent; or c) FTA-ABS-CSF with result “reactive” or “positive” or equivalent.

Appendix Methods: Algorithm development

We used Super Learning to evaluate 42 candidate algorithms for HIV prediction. Super Learning uses cross-validation to select an optimal combination of predictions from across multiple candidate algorithms or to identify the single best algorithm.[1,2] The candidate algorithms investigated by Super Learner, known as the Super Learner library, included logistic regression models[3] and machine learning algorithms: least absolute shrinkage and selection operator (LASSO),[4] ridge regression,[4] random forest,[5] support vector machine,[6] and neural networks.[7]

Risk prediction algorithms can have better predictive performance when cases and controls are more equally represented in the data, so algorithms were also applied to subsets of data where the ratio of controls to cases was approximately 20:1 or 10:1. Algorithms were also applied to restricted datasets containing a subset of 23 predictors that were pre-selected using clinical judgement, to assess whether inclusion of fewer covariates could decrease overfitting. These 23 pre-selected predictors are denoted in Appendix Table 1. Appendix Table 2 describes the candidate prediction algorithms in the Super Learner library, including numbers of controls per case and whether algorithms used pre-selected predictors or all predictors.

Analyses were run in the R statistical programming environment.[2] The following additional R packages that were used are available online on the Comprehensive R Archive Network (CRAN) website at http://cran.r-project.org/: Super Learner v 2.0-21[2]; glmnet v2.0-5[4;] randomForest v4.6-12[5]; e1071 v1.6-7[6]; nnet v7.3-12[7].

Appendix Table 2. Candidate prediction algorithms in the Super Learner library.

Prediction algorithm	Controls per case (approximate) n	Eligible covariates	Notes[a]
Logistic regression	50	All	Weighted
	50	Pre-Selected	
	50	All	--
	50	Pre-selected	
	20	All	Weighted
	20	Pre-selected	
	20	All	--
	20	Pre-Selected	
	10	All	Weighted
	10	Pre-Selected	
	10	All	--
	10	Pre-Selected	
	50	All	Stepwise backward selection
	50	Pre-Selected	
LASSO	20	All	Deviance loss
	20	All	AUC loss
	20	Pre-Selected	Deviance Loss
	20	Pre-Selected	AUC loss
	10	All	Deviance loss
	10	All	AUC loss
	10	Pre-Selected	Deviance loss
	10	Pre-Selected	AUC loss
Ridge regression	20	All	Deviance loss
	20	All	AUC loss
	20	Pre-Selected	Deviance loss
	20	Pre-Selected	AUC loss
	10	All	Deviance loss
	10	All	AUC loss
	10	Pre-Selected	Deviance loss
	10	Pre-Selected	AUC loss
Random forest	50	All	10,000 trees, 1/3 of covariates sampled per split
	50	Pre-Selected	
	20	All	
	20	Pre-Selected	
	10	All	
	10	Pre-Selected	
Support vector machine	50	All	Tuning parameters chosen by cross validation
	50	Pre-Selected	
Neural network	20	Pre-Selected	10 nodes in 1 hidden layer
	20	Pre-Selected	5 nodes in 1 hidden layer
	10	Pre-Selected	10 nodes in 1 hidden layer
	10	Pre-Selected	5 nodes in 1 hidden layer

LASSO, least absolute shrinkage and selection operator; AUC, area under the receiver-operating curve.

[a]Risk scores were re-scaled to account for undersampling of controls, except for weighted logistic regression algorithms.

Appendix Table 3. Demographic characteristics of patients with incident HIV infection and PrEP use in the three study cohorts.

Characteristic	Atrius Health 2007-2015		Atrius Health 2016		Fenway Health 2011-2016	
	Incident HIV infection (n=150)	PrEP use (n=90)	Incident HIV infection (n=16)	PrEP use (n=128)	Incident HIV infection (n=423)	PrEP use (n=1813)
Age (years),[a] mean (SD)	40 (12)	38 (10)	34 (12)	38 (11)	37 (11)	36 (10)
Male, n (%)	120 (80)	87 (97)	15 (94)	128 (100)	395 (93)	1764 (97)
Race, n (%)						
White	55 (37)	66 (73)	8 (50)	75 (59)	243 (57)	1375 (76)
Black or African-American	51 (34)	6 (7)	3 (19)	11 (9)	57 (13)	95 (5)
American Indian or Alaska Native	0 (0)	0 (0)	0 (0)	0 (0)	2 (0.5)	3 (0·2)
Asian	5 (3)	1 (1)	1 (6)	10 (8)	27 (6)	84 (5)
Native Hawaiian and Other Pacific Islander	1 (1)	0 (0)	0 (0)	0 (0)	2 (0.5)	9 (0·5)
Other	26 (17)	13 (14)	3 (19)	21 (16)	44 (10)	151 (8)
Hispanic or Latino, n (%)	12 (8)	6 (7)	1 (6)	11 (9)	48 (11)	96 (5)

PrEP, preexposure prophylaxis.

[a]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 electronic health records data element.

Appendix Table 4. Performance of LASSO algorithm for detecting incident HIV infection in prospective validation and external validation cohorts.[a]

Percentile of HIV risk score in development cohort used to define test positivity	Risk score (out of 100,000)	Atrius Health 2016				Fenway Health			
		Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)	Sensitivity (%)	Specificity (%)	Positive predictive value (%)	Negative predictive value (%)
10%	0	100	54·2	0	100	100	0	1·5	100
20%	0	100	54·2	0	100	100	0	1·5	100
30%	0	100	54·2	0	100	100	0	1·5	100
40%	0	100	54·2	0	100	100	0	1·5	100
50%	0	100	54·2	0	100	100	0	1·5	100
60%	0	100	54·2	0	100	100	0	1·5	100
70%	1	100	67·6	0	100	100	2·0	1·5	100
80%	2	100	75·8	0	100	98·1	26·8	1·9	99·9
90%	8	93·8	91·0	0	100	91·3	44·2	2·4	99·7
91%	9	93·8	91·1	0	100	91·0	45·3	2·4	99·7
92%	10	81·2	92·4	0	100	87·5	48·2	2·4	99·6
93%	11	81·2	93·7	0	100	85·6	51·2	2·5	99·6
94%	12	75·0	94·7	0	100	83·0	54·6	2·6	99·5
95%	13	62·5	95·4	0	100	80·4	59·1	2·8	99·5
96%	15	50·0	96·5	0	100	53·2	75·7	3·1	99·1
97%	18	50·0	97·2	0·1	100	48·9	79·1	3·3	99·1
98%	25	37·5	98·2	0·1	100	46·3	84·7	4·3	99·1
99%	32	25·0	99·0	0·1	100	31·0	92·0	5·4	98·9

LASSO, least absolute shrinkage and selection operator.

[a]Includes patients at risk for acquiring HIV who were not on or initiating PrEP and not previously diagnosed with HIV (n= 536,400 at Atrius in 2016, and n = 29,125 unique patients seen at Fenway during 2011-2016).

Appendix References

1. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol 2007; 6: Article 25.
2. Polley E, LeDell E, van Der Laan M. SuperLearner: Super Learner Prediction. R package version 2.0-19. 2018. Accessed at: https://CRAN.R-project.org/package=SuperLearner on March 22, 2019.
3. R Development Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria, 2016. Accessed at: https://www.R-project.org on January 18, 2019.
4. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw 2010; 33(1): 1-22.
5. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 2(3):18–22. Accessed at: https://cran.r-project.org/doc/Rnews/ on January 18, 2019.
6. Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. 2017. R package e1071 version 1.6-8. Accessed at: http://CRAN.R-project.org/package=e1071 on March 22, 2019.
7. Venables WN, Ripley BD. Modern applied statistics with S. 4th ed. New York: Springer; 2002.