# Using Machine Learning to Identify Patients at Risk of Acquiring HIV in an Urban Health System

*Arun Kumar Nethi, MS,[a] Albert George Karam, MS, MBA,[a] Kristin S. Alvarez, Pharm D, BCPS,[b] Amneris Esther Luque, MD,[c] Ank E. Nijhawan, MD, MPH, MSCS,[c] Emily Adhikari, MD,[d] and Helen Lynne King, MD[c]*

**Background:** Effective measures exist to prevent the spread of HIV. However, the identification of patients who are candidates for these measures can be a challenge. A machine learning model to predict risk for HIV may enhance patient selection for proactive outreach.

**Setting:** Using data from the electronic health record at Parkland Health, 1 of the largest public healthcare systems in the country, a machine learning model is created to predict incident HIV cases. The study cohort includes any patient aged 16 or older from 2015 to 2019 (n = 458,893).

**Methods:** Implementing a 70:30 ratio random split of the data into training and validation sets with an incident rate <0.08% and stratified by incidence of HIV, the model is evaluated using a k-fold cross-validated (k = 5) area under the receiver operating characteristic curve leveraging Light Gradient Boosting Machine Algorithm, an ensemble classifier.

**Results:** The light gradient boosting machine produces the strongest predictive power to identify good candidates for HIV PrEP. A gradient boosting classifier produced the best result with an AUC of 0.88 (95% confidence interval: 0.86 to 0.89) on the training set and 0.85 (95% confidence interval: 0.81 to 0.89) on the validation set for a sensitivity of 77.8% and specificity of 75.1%.

**Conclusions:** A gradient boosting model using electronic health record data can be used to identify patients at risk of acquiring HIV and implemented in the clinical setting to build outreach for preventative interventions.

**Key Words:** HIV predictive model, machine learning for HIV risk, HIV risk model, HIV prevention

(*J Acquir Immune Defic Syndr* 2024;97:40–47)

## BACKGROUND

Despite improvements in morbidity and mortality associated with HIV due to antiretroviral therapy, the incidence of HIV has only modestly decreased, with a 9% decrease between 2015 and 2019 and a total of 36,136 cases in 2021.[1] The US Preventative Services Task Force recommends preexposure prophylaxis (PrEP) as a best practice to reduce the risk of acquiring HIV.[2,3] The US Department of Health and Human Services Ending the HIV Epidemic in the US (EHE) initiative was launched in 2019 with a goal of reducing new HIV infections in the United States by 90% by 2030. The EHE is targeting geographic areas with a disproportionate burden of incident HIV infections and working to scale up effective prevention services. However, identification and engagement of patients with an increased likelihood of HIV is often a challenge in the clinical setting and remains a logistical barrier to reaching the EHE goals. While studies surveying primary care providers have suggested increased awareness of PrEP, remaining barriers to implementing PrEP into primary care practice includes lack of familiarity with indications for PrEP and time constraints.[4,5] The need to assure privacy and confidentiality and ask behavior-related questions may be seen as time consuming and may discourage busy clinicians from engaging in HIV risk assessment.

Machine learning models have been successful in addressing process improvement challenges within the

Received for publication January 30, 2023; accepted February 13, 2024.

From the [a]PCCI, Dallas, TX; [b]Center of Innovation and Value at Parkland Health, Dallas, TX; [c]Department of Internal Medicine, Division of Infectious Diseases and Geographic Medicine, University of Texas Southwestern Medical Center, Dallas, TX; and [d]Department of Obstetrics & Gynecology, Division of Maternal Fetal Medicine, University of Texas Southwestern Medical Center, Dallas, TX.

Funding provided by Parkland Health (PH) and Parkland Center for Clinical Innovation (PCCI).

A.N. and H.L.K. receive research funding from Gilead Sciences. Authors have no conflict of interest to declare. In addition, the abovementioned contributors in the acknowledgment received no compensation for their work other than their usual salary and have no conflicts of interest relevant to this article.

A sincere thanks to all the contributions from the following teams or individuals from Parkland Center for Clinical Innovation: George Oliver (MD), Manjula Julka (MD), Aida Kreho, and Venkatraghavan Sundaram for feedback during the creation of the model. The authors would also like to acknowledge the support and cooperation received from the following individuals at Parkland Health, Dallas, TX: The Parkland Community Integrated Health department for the management of the community health needs assessment, including the work of Swati Singh (Program Manager), Yoomi Kwon (APP), Cornelius Brewer-Bell (RN), Lametra Leon (RN), Traneika Turner (LCSW), Crystal Curtis, and Christine Joll who work daily to ensure that patients are screened routinely for HIV.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.corneajrnl.com).

Correspondence to: Arun Kumar Nethi, MS, PCCI, 8435 N Stemmons Fwy, #1150, Dallas, TX 75247, (e-mail: Arun.Nethi@pccinnovation.org).

clinical workflow.[6,7] For example, for sepsis, early identification and diagnosis are imperative to mitigate the risk of severe health outcomes, and several machine learning models have demonstrated usefulness in early identification.[7]

HIV predictive models using electronic health record (EHR) data have been described and are a promising tool to assist in the implementation of HIV prevention efforts.[8–14] Such a predictive model built into the EHR may help providers identify individuals eligible for PrEP.

## SETTING

Prior studies have utilized patient data from the EHR to develop prediction models for incident HIV.[8–14] These models have the potential to be used clinically to identify patients with an increased likelihood of HIV before acquisition and support prevention interventions. However, there are very few published models that (1) include data from the southern areas of the United States and uninsured populations,[13] (2) are able to identify cases in all subpopulations by sex at birth[15,16] effectively without multiple specialized models for each subpopulation, and (3) utilize EHR data alone for ease and scalability in similar healthcare settings. Prior models address some of these criteria but not all. The development of predictive models that meet these criteria is imperative to effectively (1) address the needs of vulnerable populations and (2) implement them in hospital settings with limited resources.

In this study, a prediction model is developed for identification of HIV cases at Parkland Health (PH), one of the largest safety-net hospitals in the country, which serves a large uninsured population in Dallas County, Texas. Dallas County was identified by the EHE as having a disproportionate burden of incident HIV and is a target jurisdiction of the EHE initiative.

## METHODS

### Data

At PH, EHR data were extracted from Epic (Epic Systems Corporation, Verona, WI) for any patient aged 16 or older from 2015 to 2019. There were 458,893 unique patients with at least 1 inpatient, emergency, or outpatient visit within the period. The incident HIV population is defined using information available from PH's HIV registry, laboratory results, diagnosis information, and previous HIV screening data followed by chart reviews for any incident case between 2015 and 2019. Each newly positive HIV screening test in the health system is reviewed by clinical staff who perform direct outreach to the patient and the local health department to determine and document whether the result represents an incident versus a known HIV diagnosis. Any patient who had a visit within the defined timeline but was diagnosed with HIV before 2015 was excluded from the analysis. Overall, the incidence rate of HIV is approximately 0.08% (or 80 per 100,000) in the cohort. Data on 60 different variables encompassing demographic, social history, laboratory test results, medical diagnoses, medication history, and hospital utilization history for the population were included based on published models[8–10] and input from clinical experts. Furthermore, owing to the rarity of certain variables, some variables are grouped into related categories (eg, stomach cancer, skin cancer, digestive cancer, etc. were grouped into a variable named non-AIDS-defining cancers). Table 1, Supplemental Digital Content (see http://links.lww.com/QAI/C311) provides a description of several variables considered for the predictive model.

### Training and Validation Data Set

The data set created was then split randomly in the ratio 70:30 to create training data with 70% patient records and validation data with 30% patient records. To ensure that the incident rate remained similar in both splits, the data split was stratified by the outcome variable of incident HIV cases. This resulted in a training data set of 322,142 unique patients and a validation data set of 136,751 unique patients. Both data sets had a similar prevalence rate of $<0.08\%$.

### Variable Transformations

Univariate analysis was performed to understand the distributions of the variables. To handle varied scale ranges, all continuous variables were transformed into categorical variables, either by binning them into finite categories or by transforming them into binary variables based on training data availability. For example, the univariate analysis showed the presence of a higher HIV rate in the population below the age of 47. Hence, age was converted to a binary variable, which is flagged as 1 if the patient is $<47$ years old and 0 if the patient is $\geq 47$ years old. For continuous variables, such as number of visits, the variable was binned into 4 categories (0 visits [no prior visit history], 1–10 visits, 11–30 visits, and $\geq 31$ visits) based on the distribution of the variable. Similar approaches were used for all variables, as applicable, to transform them into categorical data.

Our data did have missing data which were handled depending on the variable type or missing counts. The following approaches were employed: (1) In case of missing data (eg, contraceptive usage), a binary variable was created if no information was available for a particular patient to indicate missing information for the variable; (2) In case of variables related to laboratory test results (eg, chlamydia positivity), patients were labeled "1" if positive previously and were labeled "0" if either negative previously or if no information is available in the EHR; (3) In other cases, where missing values were few or negligible (eg, sex, which had $<0.2\%$ missing), the variable was converted to binary (eg, patient was labeled "1" if male with all the non-male population and missing values labeled "0").

The variables were then checked for multicollinearity using the variance inflation factor (VIF). VIF explains the strength of correlated independent (or predictor) variables. Handling correlated independent variables is important to avoid any inflation in the variance of the model with highly correlated variables. For example, an independent variable

"Smoking Frequency—Annually" could be highly correlated with another independent variable "Smoking—Number of times per week." Usage of both highly correlated variables can result in skewed or misleading results, particularly in linear models like Logistic Regression. VIF with a value of ≥10 for a variable implies that the variable is highly correlated. Hence, those features with VIF ≥10 were excluded from the variables list. This resulted in 29 variables from the original list of potential variables (see Table 1, Supplemental Digital Content, http://links.lww.com/QAI/C311).

## Algorithm and Training

An initial application of the stratified K-fold cross-validation (k = 5) technique for model training is implemented. Stratified K-fold cross-validation splits the training data into k splits stratified by the outcome variable to ensure the incident rate is similar in all the data splits and then trains the model on k–1 splits and validates the model on the kth split. This method iteratively trains the model on all the data splits available until each split is used as a validation set. Then, the average of performances is considered for optimization and generalization.

Since it is unknown which machine learning algorithm can learn from the input data better before any model training and evaluation, different algorithms can be trained, and their performance can be compared to identify the best-performing model. While linear models like Logistic Regression are interpretable, ensemble models can be useful to boost predictive performance and to avoid issues like overfitting.

After training several models using machine learning algorithms including Logistic Regression, Random Forest, XGBoost, and Balanced Bagging Classifier (Fig. 1); Light Gradient Boosting Machine (LGBM) Algorithm,[17] which is an ensemble classifier, was the best-performing model based on the evaluation metrics such as sensitivity and specificity. A Decision Tree classifier is used as the base estimator in the ensemble. Boosting combines a set of weak learners (in this study, the weak learner used is a Decision Tree classifier) that are trained sequentially, with each successor weak learner improving from the errors of the predecessor weak learner to result in a strong boosting classifier.

While all the variables were used for training initially, variable selection was done using the relative variable importance derived from the ensemble classifier. Relative variable importance is calculated by the number of times a feature is used to split the data in the model. At each split, a variable usage is determined depending on the maximization of homogeneity of the subsets created by the split.[17]

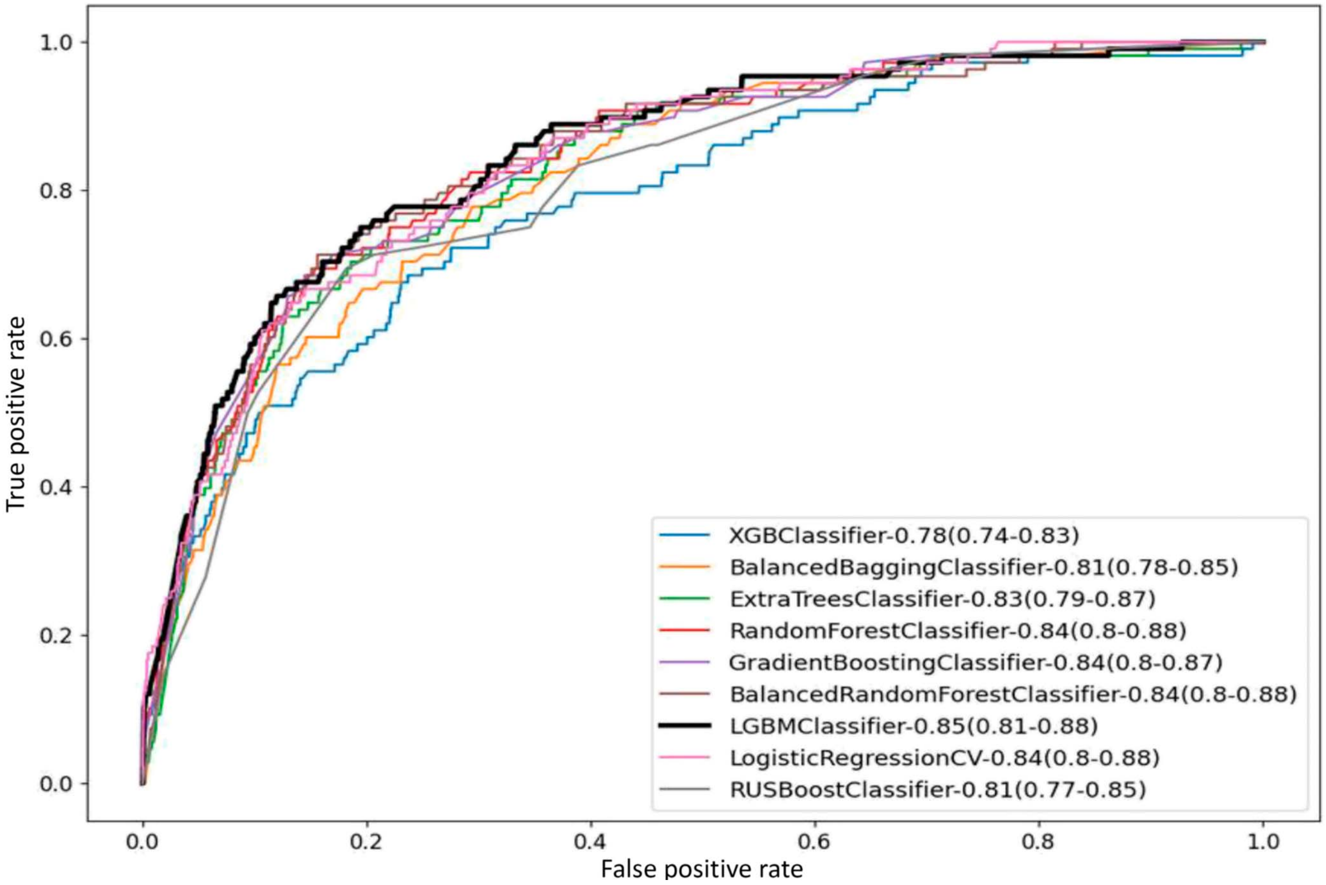

**FIGURE 1.** ROC curves for the candidate models (legend displays model name—corresponding AUC with CIs).

**TABLE 1.** Performance of LGBM Model on Validation Cohort

| Top Percent of Scores (%) | Sensitivity (%) | Positive Predictive Value (%) | Specificity (%) | Negative Predictive Value (%) | Risk Score |
|---|---|---|---|---|---|
| 1 | 13 | 1.5 | 99.3 | 99.9 | 0.8 |
| 3 | 20.4 | 1 | 98.4 | 99.9 | 0.78 |
| 5 | 41.7 | 0.7 | 95 | 100 | 0.73 |
| 10 | 57.4 | 0.5 | 90 | 100 | 0.63 |
| 15 | 66.7 | 0.4 | 85.7 | 100 | 0.56 |
| 20 | 70.4 | 0.3 | 80.3 | 100 | 0.52 |
| 25 | 77.8 | 0.2 | 75.1 | 100 | 0.49 |
| 30 | 82.4 | 0.2 | 70.1 | 100 | 0.46 |
| 40 | 88.9 | 0.2 | 60 | 100 | 0.38 |
| 50 | 92.6 | 0.1 | 50.2 | 100 | 0.3 |
| 60 | 94.4 | 0.1 | 40.7 | 100 | 0.26 |
| 70 | 98.1 | 0.1 | 30 | 100 | 0.21 |
| 80 | 98.1 | 0.1 | 20.1 | 100 | 0.17 |
| 90 | 99.1 | 0.1 | 10 | 100 | 0.12 |
| 100 | 100 | 0.1 | 0 | 100 | 0.08 |

Given the low prevalence rate and high-class imbalance of the outcome variable in the data set, to avoid any bias toward the majority class, the weights for the classes were adjusted inversely proportional to the class frequencies. Other sampling strategies such as SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic) to oversample the minority classes were explored as potential options that resulted in lower accuracy results.

## Evaluation

Evaluation metrics such as area under the receiver operator characteristic (ROC) curve (AUC), sensitivity, specificity, and positive and negative predictive values with 95% confidence intervals for the validation data set identified the best possible model. An evaluation, found in Table 1, of the best model at different risk thresholds was applied to better understand the model's performance in the identification of incident HIV cases.

# RESULTS

## Model Results

The cohort of 458,893 PH patients was split into a training set of 322,142 patients (consisting of 256 incident HIV cases) and a validation set of 136,751 patients (consisting of 108 incident HIV cases). In the training data, 78 patients of 256 incident HIV cases were women, and 27 of 108 incident HIV cases were women in the validation data. Table 2 shows that the randomly split training and validation sets based on stratification of the outcome variable resulted in similar demographic distributions across age, race, ethnicity, and marital status.

The LGBM model had the best discrimination with an AUC of 0.88 (95% confidence interval [CI]: 0.86 to 0.89) on the training set and when validated using the randomly split validation data set, it resulted in an AUC of 0.85 (95% CI: 0.81 to 0.89). Other models trained had AUCs in the range of 0.785–0.848 using the validation cohort (Fig. 1). While the other models were close enough in performance, the decision to choose a boosting model was based on the most optimal performance using metrics such as sensitivity and positive

**TABLE 2.** Demographics Table Comparing the Training and Validation Cohorts

| Variable | Training (n = 322,142) | Validation (n = 136,751) |
|---|---|---|
| Age | 42.5 (15.9) | 42.5 (15.9) |
| Birth sex | | |
| Female | 209,410 (65.0) | 89,417 (65.4) |
| Male | 112,676 (34.9) | 47,307 (34.6) |
| Other/Unknown | 56 (0.02) | 27 (0.02) |
| Ethnicity and first race | | |
| Hispanic | | |
| White | 168,772 (52.7) | 71,751 (52.5) |
| Black | 791 (0.3) | 316 (0.2) |
| Others | 2360 (0.7) | 1046 (0.7) |
| Non-Hispanic | | |
| White | 43,708 (13.6) | 18,389 (13.5) |
| Black | 88,301 (27.4) | 37,850 (27.7) |
| Others | 13,171 (4.1) | 5772 (4.2) |
| Unknown | 4039 (1.2) | 1627 (1.2) |
| Marital status | | |
| Common law | 19,658 (6.1) | 8366 (6.1) |
| Married | 100,449 (31.1) | 42,529 (31.1) |
| Single | 149,401 (46.4) | 63,394 (46.4) |
| Other/unknown | 52,634 (16.3) | 22,462 (16.4) |
| Insurance coverage type | | |
| Uninsured | 269,517 (83.68) | 114,317 (83.6) |
| Medicare/Medicaid | 40,550 (12.59) | 17,369 (12.7) |
| Commercial | 11,450 (3.56) | 4784 (3.5) |
| Unknown | 625 (0.19) | 281 (0.21) |
| Incident HIV | 256 (0.1) | 108 (0.1) |

Data are represented as mean (SD) for age and counts (percentage) for other variables.

predictive value (PPV) along with the AUC. For example, the use of Logistic Regression in the validation cohort could have resulted in an additional 3380 false positives and 3 additional false negatives as compared with the LGBM classifier. The LGBM classification model predicted 77.8% (84 of 108) of the total HIV cases in the validation set to be of high risk. The 84 true positives include 91.4% (74 of 81) of male cases and 37% (10 of 27) of female cases; 73% (54 of 74) of non-Hispanic cases and 85% (29 of 34) of Hispanic cases; 69.3% (43 of 62) of Race-African American cases, 86.3% (38 of 44) of Race-White cases; 80% (76 of 95) of uninsured cases, 100% (4 of 4) of cases with commercial insurance, and 33.3% (3 of 9) of cases with Medicare/Medicaid insurance.

The validation cohort included 12 patients who were men who had sex with men among the 108 HIV cases, and the model predicted high risk for 100% (12 of 12) of them.

The final model selected has the top 26 variables by variable importance. Number of variables, n = 26, was chosen after evaluating several values of n on the best combination of performance metrics such as AUC, sensitivity, and

**TABLE 3.** HIV Risk Predictors in the Development Data Set

| Variable | Non-HIV (n = 458,529) | Non-HIV Data Availability | HIV (n = 364) | HIV Data Availability |
|---|---|---|---|---|
| Age | 42.483 (15.9) | 458,529 (100) | 34.81 (12.1) | 364 (100) |
| Use of contraceptives | | 92,268 (20.1) | | 51 (14) |
| Condom | 26,248 (5.7) | | 33 (9.1) | |
| Others | 66,020 (14.4) | | 18 (4.9) | |
| Demographics | | | | |
| Marital status—single | 212,501 (46.4) | 452,639 (98.7) | 294 (80.8) | 354 (97.2) |
| Birth sex—male | 159,724 (34.9) | 458,529 (99.8) | 259 (71.2) | 364 (100) |
| Language—English | 284,756 (62.1) | 386,478 (84.3) | 303 (83.2) | 364 (100) |
| Tobacco use | 148,543 (32.4) | 364,245 (79.3) | 193 (53.0) | 326 (89.5) |
| Ethnicity and race | | 452,867 (98.7) | | 360 (98.9) |
| Hispanic | | | | |
| White | 241,400 (52.7) | | 123 (33.8) | |
| Black | 1106 (0.2) | | 1 (0.3) | |
| Others | 3399 (0.7) | | 7 (1.9) | |
| Non-Hispanic | | | | |
| White | 62,050 (13.5) | | 47 (12.9) | |
| Black | 125,972 (27.5) | | 179 (49.2) | |
| Others | 18,940 (4.1) | | 3 (0.8) | |
| Unknown | 5662 (1.2) | | 4 (1.1) | |
| Social history | | | | |
| Men who have sex with men (MSM) | 1324 (0.3) | 1324 (0.3) | 31 (8.5) | 31 (8.5) |
| Sexually active | 163,496 (35.7) | 365,388 (79.7) | 115 (31.6) | 212 (58.2) |
| Tobacco use | 73,931 (16.1) | 185,452 (40.4) | 147 (40.4) | 340 (93.4) |
| Alcohol use | 101,111 (22.1) | 361,882 (78.9) | 135 (37.1) | 358 (98.3) |
| Drug use | 35,099 (7.7) | 96,331 (21) | 92 (25.3) | 334 (91.7) |
| Any (+) chlamydia screening | 10,050 (2.2) | 10,050 (2.2) | 10 (2.8) | 10 (2.8) |
| Any prior HIV RNA tests | 164,360 (35.9) | 164,360 (35.9) | 135 (37.1) | 135 (37.1) |
| Any amphetamine-positive tests | 4672 (1.0) | 4672 (1.0) | 24 (6.6) | 24 (6.6) |
| Hospital utilization | | 458,529 (100) | | 364 (100) |
| Inpatient admissions in the past 2 yrs | 74,224 (16.2) | | 44 (12.1) | |
| Total health system visits | | | | |
| <1 | 2055 (0.5) | | 0 | |
| 1 to 10 | 108,735 (23.7) | | 176 (48.4) | |
| 11 to 30 | 113,394 (24.7) | | 88 (24.2) | |
| ≥31 | 234,345 (51.1) | | 100 (27.5) | |
| Exposure to venereal disease (ICD 10 code) | 8984 (1.9) | 8984 (1.9) | 25 (6.8) | 25 (6.8) |
| ICD-10 code indicating risk | 3274 (0.7) | 3274 (0.7) | 13 (3.6) | 13 (3.6) |
| History of hepatitis C | 13,771 (3.0) | 13,771 (3.0) | 20 (5.5) | 20 (5.5) |
| History of syphilis | 2267 (0.49) | 2267 (0.49) | 15 (4.12) | 15 (4.12) |
| History of lower respiratory tract infection | 71,281 (15.6) | 71,281 (15.6) | 49 (13.5) | 49 (13.5) |
| History of non-HIV infections | 38,346 (8.4) | 38,346 (8.4) | 23 (6.3) | 23 (6.3) |
| History of non-AIDS-defining cancers | 5277 (1.2) | 5277 (1.2) | 2 (0.6) | 2 (0.6) |

Data are represented as mean (SD) for age and counts (percentage) for other variables. Data availability is represented as counts (percentage) for all variables.

specificity. Table 3 presents the comparison of the variables retained in the model for the patients without HIV and with HIV. Figure 2 shows the importance of the variables used in the model. The top 5 predictors in the 26 variables used in the model were as follows: marital status being single or unmarried; age being less than 47; sex at birth being male; total visits in the range of 1–10; and ethnicity being non-Hispanic.

Table 1 presents the sensitivity, PPV, specificity, and negative predictive value for the model at different thresholds of risk in the validation cohort. Positive predictive values remain low, while negative predictive values remain high always.

## Further Discussion Results and Limitations

Overall, the predictive model resulted in an AUC of 0.85 (95% CI: 0.81 to 0.89) with the ability to segment the patient population between HIV and non-HIV. Despite the low incidence rate overall (<0.08%), the model was able to identify 77.8% of the HIV cases at the default threshold of 0.50, flagging patients with a score equal to or above 0.50 as high risk and below 0.50 as low risk. This demonstrates that machine learning models could be used for predicting and classifying the risk of HIV using available EHR data.

Similar to other studies and consistent with local trends in HIV epidemiology, in the cohort, the incidence rate based on sex assigned at birth was higher in men compared with women.[8,9] Prediction tools that can accurately identify female patients with an increased likelihood of HIV and who are eligible for PrEP are needed to improve the implementation of prevention interventions in this population. The Centers for Disease Control and Prevention (CDC) estimates that 6.5% of those assigned female at birth with indications for PrEP were prescribed PrEP in 2018 compared with 20.9% of those assigned male at birth.[18] While Krakower et al and Marcus et al models were unable to identify risk in female patients with HIV, our model was able to identify 10 of 27 women with incident HIV as at increased likelihood of HIV in the unseen validation data set.[8,9] Although the Friedman et al[12] model was trained only on a female population, our approach of using a unified model that can identify different sub-populations of HIV risk allows for a single workflow. This prediction model may be used to support clinicians in identifying female patients with an increased likelihood of acquiring HIV and linking these patients to preventative services.

Our model also expanded on the capabilities of prior models with the inclusion of a population in the southern United States, where a disproportionate burden of incident HIV is diagnosed.[1] While the Burns et al[13] model also was able to predict the likelihood of HIV for a population in the southern United States, this model relies on variable data related to geographical incidence rates of HIV and, social determinants from external data sources. An advantage to our model is that it is built solely based on the availability of data

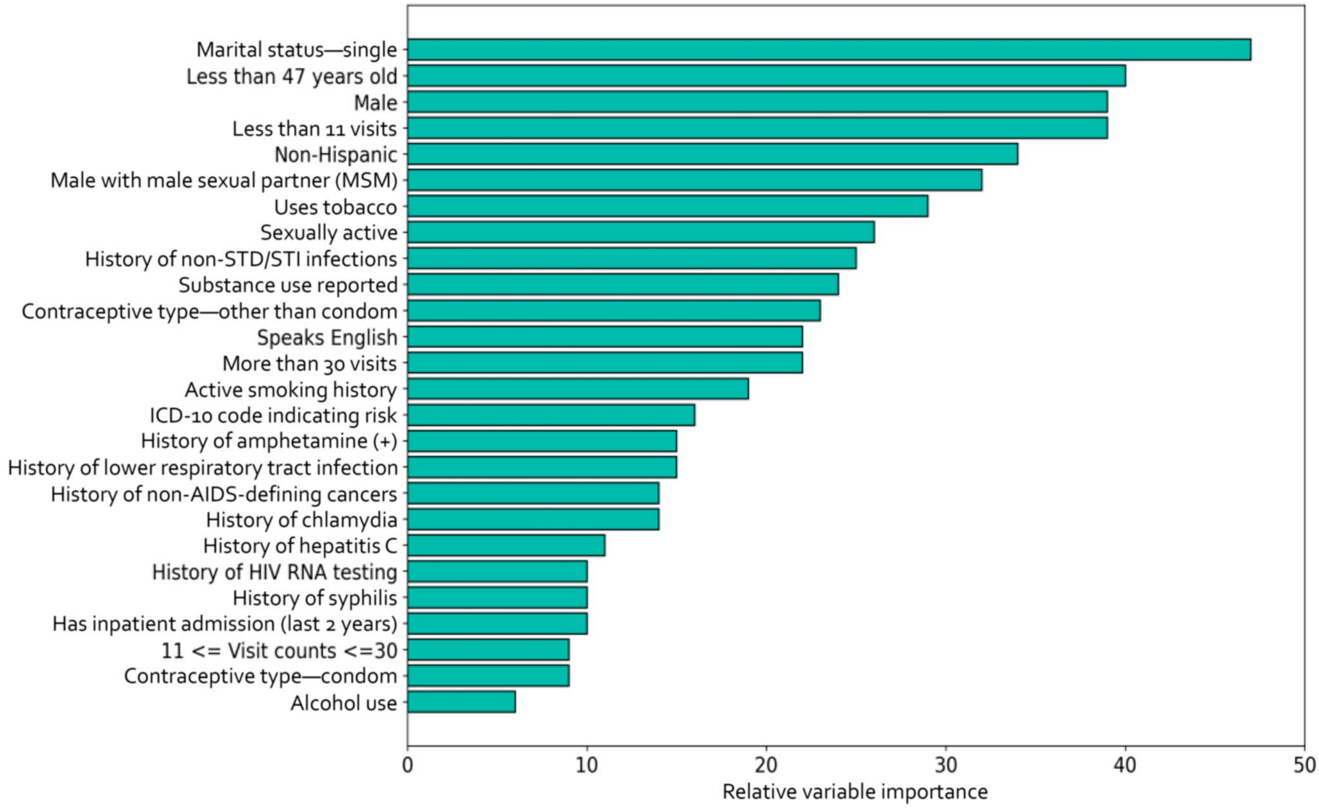

**FIGURE 2.** Relative variable importance.

from the EHR allowing the model's application in settings with EHR data alone. In addition, our cohort included a population that is 54% Hispanic and 27% non-Hispanic Black, whereas other models were based on populations that were not as diverse.

There were several other notable differences in our model from other published models. For instance, our model is trained on a population that is largely uninsured, as opposed to the May et al[11] model, which utilized data from patients with private insurance. Our training and validation data sets included 83.7% and 83.6% of patients with no insurance coverage, 12.6% and 12.7% of patients with Medicare/Medicaid insurance, and 3.6% and 3.5% of patients with commercial insurance. PH, a safety-net hospital predominantly provides health care for underserved populations across Dallas County. Our reported model outputs are very comparable to published models in predicting HIV (see Table 2, Supplemental Digital Content, http://links.lww.com/QAI/C312) with a larger potential impact given the underutilization of PrEP in the US South, among racial/ethnic minority groups and marginalized populations.

The CDC recommends that medical providers counsel all sexually active patients about PrEP and prescribe PrEP to those who are at increased likelihood of acquiring HIV. While the PPV of our EHR model is low, the goal of implementation of such a model is to cast a wide net to assist providers in identifying more patients who might be eligible for PrEP and prompt a more in-depth history and assessment of PrEP eligibility of the patients identified by the model.[19] The goal of implementation of such a model should not be to imply that all patients who have a high prediction score should be prescribed PrEP. Successful implementation of an EHR prediction model into clinical practice must be paired with provider tools to guide discussion and assessment of indications and eligibility criteria for PrEP with patients.

Prior studies indicate high acceptance by primary care providers of incorporating automated HIV prediction models into clinical practice. Providers perceived that the use of HIV predictive tools would assist in identifying PrEP-eligible patients who may otherwise be missed, standardize and destigmatize risk assessments, and serve as an educational tool to help patients visualize and perceive risk.[20,21] As models such as ours are implemented into clinical practice, it will be essential to further assess both provider and patients' perspectives on the utility of the model and identify any unintended consequences, such as concerns over privacy and model accuracy. The study has notable limitations. The predictive model was designed to predict HIV at a single center, so this exact model may not accurately predict incident HIV infections in other institutions or geographic areas due to variations in HIV epidemiologic trends. The PPV is low given the low prevalence rate in the cohort, providing an opportunity for improvement in the future. The model is also validated on data from a time before multiple interventions institutionally to increase the uptake of HIV prevention interventions, including PrEP, that may decrease the likelihood of HIV. These interventions should be included in future updates to the model to assess the impact on risk and PPV. Our study cohort might have missed HIV diagnoses by both the adjudicators and the training program.

In addition, our institutional incidence rate is higher than that of Dallas County as a whole (0.08% vs 0.03%) likely due to universal Emergency Department HIV testing protocols and a higher prevalence of HIV in our county safety-net patient population.[22] Unlike the few existing models, this model does not directly use geographical variables or prevalence rates by zip codes, though its performance was comparable to prior models that have used these variables. This geographic independence could provide an opportunity to scale and evaluate the model's usage in other hospital settings elsewhere. Finally, the ability of our EHR to detect MSM is limited and requires the use of discrete data of sex assigned at birth and sexual partner reported as male to define MSM, which could miss patients within this category.

## CONCLUSIONS

This study presents the usefulness of machine learning models that leverage EHR data to optimize patient care and outreach for populations with an increased likelihood of acquiring HIV. Furthermore, this study was conducted with data from a diverse, largely uninsured patient population residing in an EHE priority jurisdiction. Future prospective studies should evaluate different risk thresholds of the machine learning model for optimal reach to increase HIV testing and prevention interventions without compromising clinical workflow.

## ACKNOWLEDGMENTS

*Parkland Health provided partial salary support and in-kind support for the duration of the initiative activities.*

## REFERENCES

1. Centers for Disease Control and Prevention. Diagnoses of HIV infection in the United States and dependent areas, 2021. HIV surveillance report 2023; 34. Published May 23, 2023. Accessed January 22, 2024.
2. US Preventive Services Task Force. Preexposure prophylaxis for the prevention of HIV infection: US preventive services Task Force recommendation statement. *JAMA*. 2019;321:2203–2213.
3. Centers for Disease Control and Prevention. Monitoring selected national HIV prevention and care objectives by using HIV surveillance data—United States and 6 dependent areas, 2019. HIV Surveillance Supplemental Report 2021;26(No. 2). Available at: http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. https://www.cdc.gov/hiv/library/reports/hiv-surveillance/vol-32/content/special-focus-profiles.html#Women. Accessed January 20, 2023.
4. Babiarz J, Nix CD, Bowden S, et al. Insufficient PrEParation: an assessment of primary care prescribing habits and use of pre-exposure prophylaxis in patients at risk of HIV acquisition at a single medical centre. *Sex Transm Infect*. 2023;99(4):276–278.
5. Pleuhs B, Quinn KG, Walsh JL, et al. Health care provider barriers to HIV pre-exposure prophylaxis in the United States: a systematic review. *AIDS Patient Care STDS*. 2020;34:111–123.
6. Hou N, Li M, He L, et al. Predicting 30-days mortality for MIMIC-III patients with sepsis-3: a machine learning approach using XGboost. *J Transl Med*. 2020;18:462.
7. Starr AJ, Julka M, Nethi A, et al. Parkland trauma index of mortality: real-time predictive model for trauma patients. *J Orthop Trauma*. 2022; 36:280–286.
8. Marcus JL, Hurley LB, Krakower DS, et al. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modelling study. *Lancet HIV*. 2019;6:e688–e695.

9. Krakower DS, Gruber S, Hsu K, et al. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modelling study. *Lancet HIV*. 2019;6:e696–e704.
10. Ahlstrom MagnusG, Ronit Andreas, Omland LarsHaukali, et al. Algorithmic prediction of HIV status using nation-wide electronic registry data. *EClinicalMedicine*. 2019;17:100203.
11. May SB, Giordano TP, Gottlieb A. Generalizable pipeline for constructing HIV risk prediction models across electronic health record systems. *J Am Med Inform Assoc*. 2023;31(3):666–673.
12. Friedman EE, Shankaran S, Devlin SA, et al. Development of a predictive model for identifying women vulnerable to HIV in Chicago. *BMC Women's Health*. 2023;23:313.
13. Burns CM, Pung L, Witt D, et al. Development of a human immunodeficiency virus risk prediction model using electronic health record data from an academic health system in the Southern United States. *Clin Infect Dis*. 2022;76:299–306.
14. Krakower DS, Lieberman M, Marcus JL, et al. Implementing an automated prediction model to improve prescribing of HIV preexposure prophylaxis. *NEJM Catalyst*. 2023;4. https://catalyst.nejm.org/doi/abs/10.1056/CAT.23.0215. Accessed December 5, 2023.
15. Bradley ELP, Hoover KW. Improving HIV preexposure prophylaxis implementation for women: summary of key findings from a discussion series with women's HIV prevention experts. *Womens Health Issues*. 2019;29:3–7.
16. Loutfy MR, Sherr L, Sonnenberg-Schwan U, et al. Women for Positive Action. Caring for women living with HIV: gaps in the evidence. *J Int AIDS Soc*. 2013;16:18509.
17. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*. 2017. https://papers.nips.cc/paper_files/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html. Accessed June 6, 2021.
18. Centers for Disease Control and Prevention. HIV Surveillance Data Tables (early release): core indicators for monitoring the Ending the HIV Epidemic initiative (preliminary data): HIV diagnoses and linkage to HIV medical care, 2019 (reported through December 2019); and preexposure prophylaxis (PrEP)—2018, Updated HIV Surveillance Data Tables. 2020;1. Available at: http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Accessed December 7, 2022.
19. Centers for Disease Control and Prevention. US Public Health Service: Preexposure prophylaxis for the prevention of HIV infection in the United States—2021 Update: a clinical practice guideline; 2021. Available at: https://www.cdc.gov/hiv/pdf/risk/prep/cdc-hiv-prep-guidelines-2021.pdf. Accessed October 15, 2023.
20. Gilkey MB, Marcus JL, Garrell JM, et al. Using HIV risk prediction tools to identify candidates for pre-exposure prophylaxis: perspectives from patients and primary care providers. *AIDS Patient Care STDS*. 2019;33:372–378.
21. Van den Berg P, Powell VE, Wilson IB, et al. Primary care providers' perspectives on using automated HIV risk prediction models to identify potential candidates for pre-exposure prophylaxis. *AIDS Behav*. 2021;25:3651–3657.
22. Dallas County Health and Human Services. 2017 profile of HIV in Dallas county. Published August 14, 2018. Available at: https://www.dallascounty.org/Assets/uploads/docs/hhs/epistats/HIVSTIProfiles2017.pdf. Accessed January 30, 2024.