Clinical Infectious Diseases

MAJOR ARTICLE

Development of a Machine Learning Modelling Tool for Predicting HIV Incidence Using Public Health Data from a County in the Southern United States

Carlos S. Saldana, MD[1*]; Elizabeth Burkhardt, MSPH[2]; Alfred Pennisi, MA, MPH[2]; Kirsten Oliver, MPH[2]; John Olmstead, MPH[2]; David P. Holland, MD, MHS[3-4]; Jenna Gettings, DVM, MPH[2]; Daniel Mauck, MPH, PhD[2]; David Austin, BS[2], Pascale Wortley, MD[2]; Karla V. Saldana Ochoa, PhD[5].**

1. Division of Infectious Diseases, Department of Medicine, Emory University School of Medicine, Atlanta – GA, United States.; 2. Georgia Department of Public Health, Atlanta – GA, United States.; 3. Mercy Care Health Systems, Atlanta - GA, United States.; 4. Fulton County Board of Health, Atlanta GA, United States.; 5. School of Architecture, University of Florida, Gainesville - FL , United States.

Background: Recent advancements in Machine Learning (ML) have significantly improved the accuracy of models predicting HIV incidence. These models typically utilize electronic medical records and patient registries. This study aims to broaden the application of these tools by utilizing de-identified public health datasets for notifiable sexually transmitted infections (STIs) from a southern U.S. County known for high HIV incidence rates. The goal is to assess the feasibility and accuracy of ML in predicting HIV incidence, which could potentially inform and enhance public health interventions.

Methods: We analyzed two de-identified public health datasets, spanning January 2010 to December 2021, focusing on notifiable STIs. Our process involved data processing and feature extraction, including sociodemographic factors, STI cases, and social vulnerability index (SVI)

Corresponding author: Carlos S. Saldana, 341 Ponce de Leon Ave NE, Atlanta – Georgia, 30308, United States, cssalda@emory.edu.

Alternate Corresponding author: Karla V. Saldana Ochoa ksaldanaochoa@ufl.edu

DOI: 10.1093/cid/ciae100

metrics. Various ML algorithms were trained and evaluated for predicting HIV incidence, using metrics such as accuracy, precision, recall, and F1 score.

Results: The study included 85,224 individuals, with 2,027 (2.37%) newly diagnosed with HIV during the study period. The ML models demonstrated high performance in predicting HIV incidence among males and females. Influential predictive features for males included age at STI diagnosis, previous STI information, provider type, and SVI. For females, they included age, ethnicity, previous STIs information, overall SVI, and race.

Conclusions: The high accuracy of our ML models in predicting HIV incidence highlights the potential of using public health datasets for public health interventions such as tailored HIV testing and prevention. While these findings are promising, further research is needed to translate these models into practical public health applications.

Keywords: #machine learning #HIV #public health #EHE #artificial intelligence

INTRODUCTION

Despite significant advancements in the treatment and prevention of HIV, disparities remain in the implementation of these interventions, particularly in the Southern U.S. [1-3]. Fulton County, GA, is a priority jurisdiction in the Ending the HIV Epidemic (EHE) initiative due to its high HIV incidence [4]. In 2021, there were 59.6 new HIV diagnoses per 100,000 people in Fulton County, well above the national average of 11.5 [4,5]. There is a need for innovative, data-driven strategies to guide public health interventions aimed at reducing HIV rates by improving access to HIV testing and prevention.

In recent years, Artificial Intelligence (AI), particularly Machine Learning (ML) models, have been used to analyze large-scale datasets, such as electronic medical records and a variety of patient registries [6-12]. These models have been used to identify patterns and influential features associated with HIV acquisition. These tools assist and inform providers of individuals who should be prioritized for HIV testing and prevention strategies [13].

Medical records from a single healthcare setting provide an important context and inform influential features related to HIV acquisition but could fail to fully capture relevant socioeconomic features and other valuable geospatial data occurring outside of its patient population. We aim to leverage ML tools using de-identified STI/HIV public health datasets, readily available to public health officials, coupled with social vulnerability indicators from a high-incidence county in the Southern U.S. using them to predict HIV incidence. This tool could inform tailored sexual health outreach and optimize resource allocation for HIV testing and prevention.

METHODS

Study setting

Fulton County, GA, had an estimated 2021 population of 1,066,702 residents, consisting of 45% African American, 44.2% White, and 7.4% Hispanic/Latino individuals. For context, approximately 55.9% of residents held a bachelor's degree or higher. The median household income was $77,635, and the poverty rate stood at 13.7% [14].

Data sources

From January 2010 to December 2021, we collected de-identified data of individuals 13 years of age and older from two Georgia databases 1) **State Electronic Notifiable Disease Surveillance System (SendSS)**, a web-based platform that collects, manages, and analyzes data on communicable diseases such as STIs, Hepatitis C, Tuberculosis, etc. and 2) the Georgia **Electronic HIV/AIDS Reporting System (eHARS)** a browser-based system developed by the Centers for Disease Control and Prevention (CDC) in the U.S. to collect, manage, and analyze laboratory reported data specifically related to HIV. These databases collect data on demographics, diagnosis time frame, diagnosing provider type, risk behaviors, etc. We matched each individual by census tract to a **Social Vulnerability Index (SVI)** [15], a tool developed by the Agency for Toxic Substances and Disease Registry often applied in studies to assess and measure community vulnerability during epidemics, disasters, and emergencies in different populations. For our model SVI was classified using quintiles, as per tool developers, for an overall score and the four SVI themes (socioeconomic status, household composition, race/ethnicity/language, and housing/transportation). Quintiles ranged from 1 = *'very low vulnerability'*, 2 = *'low vulnerability'*, 3 = *'moderate vulnerability*', 4 = *'high 'vulnerability'* and 5 = '*very high vulnerability'.*

Dataset development

The outcome of interest was incident HIV during the study period of 2010 to 2021, defined as confirmed HIV diagnosis in eHARs. We extracted STI cases from the SendSS database, which specifically catalogs STIs cases per occurrence. Each case has both an outcome and a unique patient identifier. To ensure the accuracy of HIV diagnoses within our dataset, we applied a probabilistic matching technique to accurately cross-reference HIV diagnoses across the eHARS and SendSS databases, using key demographic characteristics. Our algorithm assigned scores to potential matches based on the similarity of these fields, with higher scores indicating a stronger likelihood of a match. We determined a positive match with a score of 67% or higher. We used this threshold as optimal balance between identifying positive matches, managing computational resources, and data limits. We excluded patients whose initial or sole record in SendSS was attributed to an HIV diagnosis and those with a documented HIV diagnosis before the study period, including those previously diagnosed out-of-state. Furthermore, subsequent STI cases after an individual's HIV event were omitted. To maintain data integrity, we also excluded cases from our analysis where the missing data exceeded a threshold of 10%. Finally, the dataset was transposed

from case-based to a patient-based model, consolidating multiple STI cases into a singular composite profile per patient, each with one or more STIs. The process of participant selection is illustrated in **Figure 1.**

Feature selection

Our study included diverse features detailed in **Table 1**, encompassing sociodemographic variables such as **age, sex assigned at birth**, **race**, and **ethnicity**. We included the **age at STI diagnosis** and compiled those in an array for individuals with multiple STI occurrences. The cumulative **non-HIV STI count** per patient was also included. Additionally, we cataloged all **previous non-HIV STIs**—including gonorrhea, chlamydia, and syphilis along with its stages—and, where multiple infections were observed in a single patient, we constructed an array to represent this. The time for **re-infection interval** was categorized by labeling re-infections as those occurring at different time intervals from the initial STI(s). **Provider type** —ranging from urgent care centers and private clinics to correctional facilities and health departments, etc.—was also organized in an array, when multiple STIs occurred. Lastly, we aligned **social vulnerability indexes** corresponding to the time of STI diagnosis to evaluate the influencing socioeconomic context.

Model development

After data pre-processing, we stratified patients by sex assigned at birth. We trained separate models for males and females, given the variety of biological and behavioral factors influencing HIV acquisition in each group. To ensure data completeness, we addressed missing values by imputing them with the mean value within the variable. Our dataset encompassed diverse features, with **numerical, categorical, and array data.** We adopted a three-fold approach to manage this heterogeneity. For **numerical data**, we implemented normalization to standardize the scale. For **categorical data**, we used one-hot encoding [16] to code the data into numerical feature vectors which were then normalized. For **array data**, we employed an ML methodology to extract numerical feature vectors using a Neural Network Autoencoder [17]. This process yielded an array of two-dimensional feature vectors for each attribute. To consolidate these two-dimensional feature vectors into a single numerical feature vector for each attribute, we employed a dimensionality reduction through T-distributed Stochastic Neighbor Embedding (T-SNE) [18]. T-SNE transformed the initial two-dimensional feature vector into a one-dimensional feature vector, which was then normalized to facilitate the analysis.

We ensured a balance between individuals **with documented HIV** and those **without** for both the training and test sets. Given that the class distribution disparity between these two groups was significant, at a 1:100 ratio, a potential bias could adversely affect the ML algorithms by neglecting the minority class, despite its crucial predictive significance. To address this class imbalance challenge, we employed resampling, a technique that involves altering the composition of the dataset [19]. Specifically, for our model, we employed 'undersampling', which entails reducing

instances from the majority class. Hence, setting a cap for examples from the majority class (without documented HIV) based on the total number of cases from the minority class (with documented HIV). Consequently, we randomly selected an equivalent number of cases from the 'without documented HIV' group to match the 'with documented HIV' establishing an unbiased representation of both classes within the training set [19]. This balance was maintained in both male and female groups, with the training set (85%) and the remaining test set (15%) reserved for validation.

Model selection and evaluation

To identify the most suitable predictive model, we adopted a comprehensive and widely recognized strategy known as the "horse race approach" [20]. This technique entails training multiple ML algorithms using the same training dataset, enabling us to evaluate their performances and select the algorithm that attains the highest predictive accuracy. We selected from a variety of established classifiers, each with distinct attributes, strengths, and limitations. The chosen classifiers included: Random Forest, Nearest Neighbors, Logistic Regression, Naive Bayes, Gradient Boosted Trees [21,22]. **Random Forest** is an ensemble of decision trees, well-suited for handling large, high-dimensional data and robust against overfitting. **Nearest Neighbors,** classifies based on proximity to the closest training examples, being simple and effective for small datasets, though it becomes slower with increasing dimensions. **Logistic Regression**, ideal for binary outcomes, easier to implement and interpret but might struggle with complex data relationships, like the multifaceted data in our dataset. **Naive Bayes**, a probabilistic classifier, applies Bayes' theorem under strong independence assumptions and is notably efficient with high-dimensional data. **Gradient Boosted Trees** incrementally build models in stages by optimizing a loss function, proving highly effective for complex datasets in both regression and classification tasks [21,22]. Through uniform training and evaluation of each model on the same training data, we establish an equitable platform for comparison. Finally, we assessed each model's performance via **accuracy** (proportion of true results), **precision** (proportion of true positive predictions in relation to the total number of positive predictions), **recall** (proportion of actual positive cases that were correctly identified by the model as positive), and **F1-score** (harmonic mean of precision and recall, providing a single measure of the model's accuracy that considers both the false positives and false negatives).

Ethical considerations

Institutional Board Review (IRB) approval was obtained through the Georgia Department of Public Health IRB Office (DPH IRB #221109). Data-sharing agreements were obtained between DPH and the University of Florida. All personal identifiers were removed from the datasets before data-sharing to maintain participant confidentiality and comply with legal standards.

RESULTS

Between 2010 and 2021, out of 132,928 STI cases recorded in SendSS, 127,169 met our inclusion criteria. When we transposed our dataset from STI-based to patient-based, identifying 85,224 unique individuals, each with one or more STI cases. Sex-assigned at birth distribution was 54% females (45,834) and 46% males (38,935), as detailed in **Table 2**. Of these, 2,027 individuals (2.37%) met our inclusion criteria and had documented incident HIV during the study period, including 1,698 males (84%) and 329 females (16%). The male training set (85%) had 1444 individuals evenly matched with 1444 randomly selected individuals 'without documented HIV' for a total of 2888 cases for this set. The pattern was maintained for the female group, which consisted of 280 in each group for a total of 560 for the training set. The remaining was used as the validation test set (15%) which for males included 508 individuals (254 with documented HIV and 254 without) and ninety-eight for females (49 with documented HIV and 49 without), as shown in **Table 3.**

On average, males were diagnosed with STIs at 28 years old and females at 24 (Range 13 to 88 and 95 respectively). In both groups, the majority were Black—63% of males and 57% of females—with a considerable portion having an unspecified race (23% of males and 32% of females). The highest number of STI cases recorded per individuals was 18 for males and 23 for females. Most experienced just a single STI episode (71% of males and 72% of females), followed by two episodes (17% in both males and females). Chlamydia was the predominant STI among females, accounting for 79% of episodes, compared to 54% in males. Conversely, gonorrhea was the most frequent in males (36%), against 19% in females. Re-infections typically occurred over a year later for both males (69%) and females (67%). When reinfections occurred within a year, they most often took place between 201-365 days from the initial STI in both genders (12%). Males were most often diagnosed at STD Clinics (22.32%) and females at private physician offices (40%), with the latter also being the second most common for males (19%) followed by hospitals for females (15%). The rarest locations for diagnoses were school based clinics for males (2%) and correctional facilities for females (2%).

Model performance and evaluation

In the analysis of various models, detailed in **Figure 2**, Gradient Boosted Trees stood out among the classifiers, achieving an 80% accuracy rate in predicting HIV incidence for both male and female groups, as detailed in **Figure 3**. For males, the model correctly identified 203 true negatives and 203 true positives. Similarly, for females, the model correctly identified 38 true negatives and 40 true positives. As depicted by the confusion matrices, the model has comparable precision and recall across both genders, although with a slightly higher error rate for females. The precision for males stands at 80%, signifying that when the model predicts an HIV case, it is correct around 80% of the time. For females, the precision is slightly lower at 78%. The recall for males, which measures the model's ability to find all actual cases of HIV was also 80%, while for females it is

slightly higher at 81%. The F1 score was equal for males and females at 80%, suggesting that the model has a balanced performance in identifying true positives and avoiding false negatives.

In terms of feature influence in prediction, for the male subset, the most influential features in order of significance, were: 1. Age at STI diagnosis, 2. Previous non-HIV STI, 3. Provider type, 4. non-HIV STI count 5. Reinfection interval, and 6. SVI theme four (housing transportation), indicating that both demographic and social vulnerability features had a notable influence. On the other hand, the predictive factors for the female subset differed slightly, emphasizing: 1. Age at STI diagnosis, 2. Ethnicity, 3. Non-HIV STI, 4. Provider type, along with Race and Overall SVI and themes two and three. These distinctions highlight the complexity and effectiveness of our model, pointing to the necessity of customizing predictive methods according to demographic specifics, as illustrated in **Figure 4.**

DISCUSSION

Our model accurately predicted HIV incidence from 2010 to 2021 leveraging de-identified public health datasets. Our models highlight sociodemographic factors influencing HIV acquisition consistent with similar trends observed throughout the Southern U.S and globally [2, 23, 24] By leveraging the capabilities of ML, 'big data', and a social context, our methodology provides a comprehensive perspective on the dynamics influencing HIV transmission. Similar approaches have been used locally and globally using EMR data and patient registries [6-12, 23,24]; however, our study is the first to utilize public health datasets in the U.S to predict HIV incidence.

Our study introduces advancements in data processing methodology, for handling multimodal data, including numerical, categorical, and arrays, as it pertains to HIV. Unlike previous approaches [7,13], which converted all features into numerical or categorical data, which could lead to information loss, our methodology effectively processes array data using established ML algorithms such as autoencoders. Additionally, we employed dimensionality reduction algorithms, transforming multidimensional data into a one-dimensional numerical feature, which can be integrated with other numerical and categorical features for training machine learning classifiers.

In our study, age at first STI diagnosis emerged as the most predictive factor for both genders, similar to a study from Sub-Saharan Africa [23]. For males, additional predictors included prior non-HIV STI occurrences and social vulnerability features, paralleling findings from other studies where socio-behavioral factors played a significant role [6,23,25]. As in other studies, these factors took precedence over race and ethnicity in males, echoing concerns about dataset biases and model fairness raised in the literature [26, 27]. For females, ethnic background and STI-related variables were predictors, aligning with insights on gender-specific risk factors seen in studies from the U.S and Africa [13, 25]. The identification of socioeconomic factors as influential predictors in these models underscores their suitability for guiding public health interventions [6]. By accurately

identifying individuals at heightened risk of HIV acquisition, these models could become valuable tools for developing strategies prioritizing individuals for HIV testing and prevention.

The strength of our study lies in its reliance on datasets routinely accessed by public health professionals, enhancing the practicality of integrating and implementing these models into existing public health frameworks. Our models demonstrate a proof of concept in effectively identifying individuals at an elevated risk of HIV acquisition within a population already at elevated risk, those with a history of STIs. This approach could offer strategic direction for enhanced HIV testing and PrEP referrals, particularly in areas with high-incidence and challenges associated to resources and or public health staffing. These tools could also guide public health officials in identifying disproportionately impacted areas, such as census tracts, for focused outreach efforts.

Our study recognizes key limitations. (i), our model's reliability hinges on the accuracy and completeness of its datasets. A significant amount of unspecified racial data may impact the model's precision, bias, and utility. Data biases, stemming from unrepresentative training data of the target population, could lead to inaccurate predictions. In this pilot phase, we acknowledge the need for data preprocessing that incorporates fairness approaches for a more human-centric model. This involves ensuring group fairness for statistical parity and individual fairness for consistent decisions [28]. (ii), our model might not accurately identify undiagnosed HIV cases, and excluding records with over 10% missing data might omit pertinent cases. (iii), it only assesses HIV risk in individuals with reportable STIs, limiting its applicability beyond this group. (iv), as our study is based in Fulton County, an urban and diverse area, it may not fully capture the broader, often rural Southern HIV epidemic, thus limiting our findings' broader applicability. Nonetheless, all Georgia jurisdictions use our model's datasets.

Future research should focus on: 1) Reducing prediction bias in models through better inclusion criteria and fairness; 2) Improving model explainability for healthcare professionals and policymakers; 3) Utilizing real-time or prospective data for real-time public health strategies; 4) Applying implementation science frameworks assessing the integration of these models in public health practice, like enhancing HIV testing and PrEP referrals; 5) Conducting qualitative research with stakeholders for best practices deploying these tools; 6) Strategies for improving data quality, including variables like gender-identity and location; 7) Prioritizing ethical considerations and community involvement for equitable, privacy-conscious model use. These steps aim to enhance the effectiveness and ethical use of predictive models in healthcare.

NOTES

Financial support. The authors have no conflict of interest to disclose related to this work.

Potential conflicts of interest. C. S. reports grants from the National Institutes of Health and Consultancy as Medical Advisor for the Office of HIV/AIDS for the Georgia Department of Public Health, and ViiV Healthcare. All other authors report no potential conflicts. All authors have submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Conflicts that the editors consider relevant to the content of the manuscript have been disclosed. EB reports PCHD grant support.

REFERENCES

Fauci AS, Redfield RR, Sigounas G, Weahkee MD, Giroir BP. Ending the HIV Epidemic: A Plan for the United States. JAMA. 2019 Mar 5;321(9):844-845. doi: 10.1001/jama.2019.1343. PMID: 30730529.

Centers for Disease Control and Prevention. HIV Surveillance Report, 2021; vol. 34. http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Published May 2023. Accessed [11/13/2023].

Doherty R, Walsh JL, Quinn KG, John SA. Association of Race and Other Social Determinants of Health With HIV Pre-Exposure Prophylaxis Use: A County-Level Analysis Using the PrEP-to-Need Ratio. AIDS Educ Prev. 2022 Jun;34(3):183-194. doi: 10.1521/aeap.2022.34.3.183. PMID: 35647866; PMCID: PMC9196948.

Bunting SR, Hunt B, Boshara A, Jacobs J, Johnson AK, Hazra A, Glick N. Examining the Correlation Between PrEP Use and Black:White Disparities in HIV Incidence in the Ending the HIV Epidemic Priority Jurisdictions. J Gen Intern Med. 2023 Feb;38(2):382-389. doi: 10.1007/s11606-022-07687-y. Epub 2022 Jun 9. PMID: 35678988; PMCID: PMC9905374.

Centers for Disease Control and Prevention. Estimated HIV incidence and prevalence in the United States, 2017–2021. HIV Surveillance Supplemental Report, 2023; 28 (No.3). http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html. Published May 2023. Accessed [11/13/2013].

Balzer LB, Havlir D V, Kamya MR, Chamie G, Charlebois ED, Clark TD, et al. Machine learning to identify persons at high-risk of human immunodeficiency virus acquisition in rural Kenya and Uganda. Clin Infect Dis. (2020) 71(9):2326–33. 10.1093/cid/ciz1096

Marcus JL, Hurley LB, Krakower DS, Alexeeff S, Silverberg MJ, Volk JE. Use of electronic health record data and machine learning to identify candidates for HIV pre-exposure prophylaxis: a modeling study. lancet HIV. (2019) 6(10):e688–95. 10.1016/S2352-3018(19)30137-7

Zheng W, Balzer L, van der Laan M, Petersen M, Collaboration S. Constrained binary classification using ensemble learning: an application to cost-efficient targeted PrEP strategies. Stat Med. (2018) 37(2):261–79. 10.1002/sim.7296

Orel E, Esra R, Estill J, Marchand-Maillet S, Merzouki A, Keiser O. Prediction of HIV status based on socio-behavioral characteristics in East and Southern Africa. PloS one. (2022) 17(3):e0264429.

Krakower DS, Gruber S, Hsu K, Menchaca JT, Maro JC, Kruskal BA, et al. Development and validation of an automated HIV prediction algorithm to identify candidates for pre-exposure prophylaxis: a modeling study. Lancet HIV. (2019) 6(10):e696–704. 10.1016/S2352-3018(19)30139-0

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. 10.1097/QAI.0000000000001580

Xu X, Ge Z, Chow EPF, Yu Z, Lee D, Wu J, et al. A machine-learning-based risk-prediction tool for HIV and sexually transmitted infections acquisition over the next 12 months. J Clin Med. (2022) 11(7):1818. 10.3390

Burns CM, Pung L, Witt D, Gao M, Sendak M, Balu S, Krakower D, Marcus JL, Okeke NL, Clement ME. 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. 2023 Jan 13;76(2):299-306. doi: 10.1093/cid/ciac775. PMID: 36125084; PMCID: PMC10202432.

United States Census Bureau. "QuickFacts Fulton County, Georgia" 2020 Census of Population and Housing, [https://www.census.gov/quickfacts/fultoncountygeorgia].

Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/ Geospatial Research, Analysis, and Services Program. CDC/ATSDR Social Vulnerability Index [2020] Database [Georgia]. https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html. Accessed on [11/13/2023].

Seger C. An investigation of categorical variable encoding techniques in machine learning: binary versus one-hot and feature hashing [Internet] [Dissertation]. 2018. (TRITA-EECS-EX). Available from: https://urn.kb.se/resolve?

Wang, W., Huang, Y., Wang, Y., & Wang, L. (2014). Generalized autoencoder: A neural network framework for dimensionality reduction. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 490-497).

Gisbrecht, A., Schulz, A., & Hammer, B. (2015). Parametric nonlinear dimensionality reduction using kernel t-SNE. Neurocomputing, 147, 71-82.

Ali, M. M. (1998). Probability models on horse-race outcomes. Journal of Applied Statistics, 25(2), 221-229

Wongvorachan T, He S, Bulut O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. Information. 2023; 14(1):54. https://doi.org/10.3390/info14010054

Allan, K. (1977). Classifiers. Language, 53(2), 285-311.

Friedman JH. Greedy function approximation: a gradient boosting machine. Annals of statistics. 2001;1189–232.

Mutai CK, McSharry PE, Ngaruye I, Musabanganji E. Use of machine learning techniques to identify HIV predictors for screening in sub-Saharan Africa. BMC Med Res Methodol. 2021 Jul 31;21(1):159. doi: 10.1186/s12874-021-01346-2. PMID: 34332540; PMCID: PMC8325403.

He J, Li J, Jiang S, Cheng W, Jiang J, Xu Y, Yang J, Zhou X, Chai C, Wu C. Application of machine learning algorithms in predicting HIV infection among men who have sex with men: Model development and validation. Front Public Health. 2022 Aug 25;10:967681. doi: 10.3389/fpubh.2022.967681. PMID: 36091522; PMCID: PMC9452878.

Birri Makota RB, Musenge E. Predicting HIV infection in the decade (2005-2015) pre-COVID-19 in Zimbabwe: A supervised classification-based machine learning approach. PLOS Digit Health. 2023 Jun 7;2(6):e0000260. doi: 10.1371/journal.pdig.0000260. PMID: 37285368; PMCID: PMC10246851.

Huang J, Galal G, Etemadi M, Vaidyanathan M.Evaluation and Mitigation of Racial Bias in Clinical Machine Learning Models: Scoping Review.JMIR Med Inform 2022;10(5):e36388.doi: 10.2196/36388

Facente SN, Lam-Hine T, Bhatta DN, Hecht J. Impact of Racial Categorization on Effect Estimates: An HIV Stigma Analysis. Am J Epidemiol. 2022 Mar 24;191(4):689-695. doi: 10.1093/aje/kwab289. PMID: 34999778.

AI Fairness 360. LF AI Incubation Project. [Internet]. [cited 2024 Jan 26]. Available from: https://ai-fairness-360.org/

TABLES AND FIGURES

Table 1. Displays data sources from Fulton County - Georgia's notifiable diseases from 2010-2021. Features are categorized by type and considerations. Data manipulation techniques, such as array development for patients with multiple STI diagnoses, are noted to clarify the approach to data integration and analysis. ID=Identification; SVI=Social vulnerability index; STI= Sexually transmitted infection; HIV=human immunodeficiency virus; CAT=categorical variable; CONT=continuous variable; ARRAY= list of categorical values

Table 1. Data sources and variables included in model development. Fulton County - Georgia 2010-2021.

Georgia's State Electronic Notifiable Disease Surveillance System (SendSS)			**Enhanced HIV/AIDS Reporting System (eHARS).**		
Variable Name	**Type**	**Considerations**	**Variable**	**Type**	**Considerations**
Patient ID	CAT	Transposed data to include a single patient with all STI cases in one record.	Matched HIV diagnosis during study period.	CAT	Used probability matching between datasets using sociodemographic information.
Sex assigned at birth	CAT	Stratified models for male and females			
Race	CAT				
Ethnicity	CAT				
Age at STI diagnosis	CAT ARRAY	For patients with multiple cases, we developed an array for the age at each STI diagnosis			
Previous Non-HIV STI	CAT ARRAY	Non-HIV STI event (i.e., Chlamydia, gonorrhea, syphilis stage). For patients with multiple cases, we developed an array for each previous non-HIV STI			

Non-HIV STI count	NUM	Count of non-HIV STI events			
Re-infection interval	CAT ARRAY	Time lapses for each STI re-infection in intervals. For patients with multiple cases, we developed an array with the interval between each STI diagnosis			
Provider type	CAT ARRAY	i.e., Hospital, clinic, correctional system, etc. For patients with multiple STI cases, we developed an array for diagnosing provider at each STI diagnosis			
Overall, SVI	CAT	Used quintiles of the first case in the dataset			
SVI Theme 1 (socioeconomic status)					
SVI Theme 2 (household composition, disability)					
SVI Theme 3 (minority status, language)					
SVI Theme 4 (housing, transportation)					

Table 2. Sociodemographic features, sexually transmitted infection (STI) data, and social vulnerability data, stratified by sex assigned at birth, and documented HIV status.

Features	**Total Patients meeting eligibility criteria** N=85,224		**Patients With Documented positive HIV** N= 2,027	
Sex assigned at birth	**Male** N=38,935	**Female** N= 45,834	**Male** N=1,698	**Female** N= 329
Age at STI mean (range)				
Age at STI mean (range)	28 (13-88)	24 (13-95)	27 (14-71)	23 (13-65)
Age at STI diagnosis				
Age at STI range	**N (%)**	**N (%)**	**N (%)**	**N (%)**
0-18	5577(9.45)	13606(20.12)	190(6.68)	93(16.40)
19-24	20479(34.70)	31982(47.29)	1112(39.07)	309(54.50)
25-34	21385(36.24)	17329(25.62)	1130(39.70)	134(23.63)
35-45	7633(12.93)	3513(5.19)	293(10.30)	22(3.88)
>50	3941(6.68)	1198(1.77)	121(4.25)	9(1.59)
Total	**59015**	**67628**	**2846**	**567**

DOI: 10.1093/cid/ciae100

Race (%)				
	N (%)	**N (%)**	**N (%)**	**N (%)**
AI/AN	236(0.61)	58(0.13)	3(0.18)	0(0.00)
Asian	0(0.00)	254(0.56)	8(0.47)	1(0.30)
Black	24646(63.50)	26372(57.67)	1384(81.51)	263(79.94)
White	4213(10.85)	3238(7.08)	194(11.43)	16(4.86)
Hawaiian/Pacific Islander	15(0.04)	25(0.05)	1(0.06)	1(0.30)
Multi-racial	139(0.36)	126(0.28)	9(0.53)	0(0.00)
Other race	515(1.33)	795(1.74)	17(1.00)	4(1.22)
Unknown	9050(23.32)	14862(32.50)	82(4.83)	44(13.37)
Total	**38814**	**45730**	**1698**	**329**
Ethnicity (%)				
	N (%)	**N (%)**	**N (%)**	**N (%)**
Hispanic	1139(2.96)	1020(2.28)	61(3.61)	12(3.67)
Non-Hispanic	26686(69.34)	2633(58.95)	1532(90.54)	254(77.68)
Unknown	10657(27.69)	17315(38.76)	99(5.85)	61(18.65)
Refused	3(0.01)	2(0.00)	0(0.00)	0(0.00)
Total	**38485**	**44674**	**1692**	**327**
Non-HIV STI count (%)				
	N (%)	**N (%)**	**N (%)**	**N (%)**
Single episode	27777(71.34)	33070(72.15)	1081(63.66)	206(62.61)
2	6852(17.60)	8002(17.46)	356(20.97)	66(20.06)
3	2289(5.88)	2666(5.82)	139(8.19)	28(8.51)
4	960(2.47)	1073(2.34)	57(3.36)	14(4.26)
>5	1057(2.71)	1023(2.23)	65(3.83)	15(4.56)
Total	**38935**	**45834**	**1698**	**329**
Previous non-HIV STIs (%)				
	N (%)	**N (%)**	**N (%)**	**N (%)**
Gonorrhea	21474(36.38)	13009(19.24)	1370(48.14)	150(26.46)

Primary Syphilis	538(0.91)	40(0.06)	55(1.93)	0(0.00)
Chlamydia	32176(54.51)	53377(78.93)	830(29.16)	396(69.84)
Syphilis, other	4838(8.20)	1202(1.78)	591(20.77)	21(3.70)
Total	**59026**	**67628**	**2846**	**567**
Reinfection interval (%)				
	N (%)	**N (%)**	**N (%)**	**N (%)**
Initial STI (s) 0-15 days	42066	48468	1893	360
Reinfection at 16 - 30 days	62(0.37)	57(0.30)	8(0.84)	2(0.97)
Reinfection at 31 - 60 days	595(3.51)	833(4.35)	36(3.78)	9(4.35)
Reinfection at 61 - 90 days	551(3.25)	763(3.98)	29(3.04)	9(4.35)
Reinfection at 91 - 120 days	577(3.40)	693(3.62)	48(5.04)	7(3.38)
Reinfection at 121 -150 days	467(2.75)	599(3.13)	41(4.30)	9(4.35)
Reinfection at 151 - 200 days	832(4.91)	949(4.95)	56(5.88)	4(1.93)
Reinfection at 201 - 365 days	2200(12.97)	2442(12.75)	161(16.89)	28(13.53)
Reinfection at > 366 days	11676(68.84)	12824(66.93)	574(60.23)	139(67.15)
Total	**16960**	**67628**	**953**	**567**
Provider type (%)				
	N (%)	**N (%)**	**N (%)**	**N (%)**
STI Clinic	13652(22.32)	7410(11.84)	684(24.03)	118(20.27)
Private Physician	11788(19.28)	24913(39.82)	499(17.53)	195(33.51)
Hospital	8757(14.32)	9408(15.04)	391(13.74)	93(15.98)
HIV Counseling Testing Site	656610.74)	1969(3.15)	551(19.36)	11(1.89)
Hospital ER/Urgent Care	4358(7.13)	4715(7.54)	150(5.27)	22(3.78)
Correctional Facility	1640(2.68)	1041(1.66)	65(2.28)	15(2.58)
Laboratory	1577(2.58)	2704(4.32)	68(2.39)	17(2.92)
School - based clinic	1524(2.49)	2255(3.60)	99(3.48)	4(0.69)
Other	11294(18.47)	8146(13.02)	339(11.91)	107(18.38)
Total	**61156**	**62561**	**2846**	**582**
Theme 1 (Socioeconomic Status)				

	N (%)	N (%)	N (%)	N (%)
1	7237(18.59)	7404(16.15)	320(18.85)	25(7.60)
2	5949(15.28)	7263(15.85)	263(15.49)	49(14.89)
3	5797(14.89)	7439(16.23)	258(15.19)	72(21.88)
4	6333(16.27)	7077(15.44)	304(17.90)	59(17.93)
5	5061(13.00)	7046(15.37)	240(14.13)	73(22.19)
Missing	8558(21.98)	9605(20.96)	313(18.43)	51(15.50)
Total	**38935**	**45834**	**1698**	**329**
Theme 2 (Household Composition)				
	N (%)	**N (%)**	**N (%)**	**N (%)**
1	7182(18.45)	6199(13.52)	387(22.79)	32(9.73)
2	7012(18.01)	7834(17.09)	278(16.37)	35(10.64)
3	5498(14.12)	7256(15.83)	250(14.72)	65(19.76)
4	5258(13.50)	7586(16.55)	234(13.78)	75(22.80)
5	5427(13.94)	7354(16.04)	236(13.90)	71(21.58)
Missing	8558(21.98)	9605(20.96)	313(18.43)	51(15.50)
Total	**38935**	**45834**	**1698**	**329**
Theme 3 (Race/Ethnicity/Language)				
	N (%)	**N (%)**	**N (%)**	**N (%)**
1	7013(18.01)	6329(13.81)	342(20.14)	38(11.55)
2	5862(15.06)	7523(16.41)	280(16.49)	70(21.28)
3	5758(14.79)	7653(16.70)	257(15.14)	54(16.41)
4	6117(15.71)	7139(15.58)	257(15.14)	53(16.11)
5	5630(14.46)	7585(16.55	249(14.66)	63(19.15)
Missing	8555(21.97)	9605(20.96)	313(18.43)	51(15.50)
Total	**38935**	**45834**	**1698**	**329**
Theme 4 (Housing/Transportation)				
	N (%)	**N (%)**	**N (%)**	**N (%)**
1	6305(16.19)	7830(17.08)	259(15.25)	46(13.98)

2	6259(16.08)	7570(16.52)	249(14.66)	54(16.41)
3	5653(14.52)	6833(14.91)	293(17.26)	68(20.67)
4	6228(16.00)	7180(15.67)	309(18.20)	47(14.29)
5	5932(15.24)	6816(14.87)	275(16.20)	63(19.15)
Missing	8558(21.98)	9605(20.96)	313(18.43)	51(15.50)
Total	**38935**	**45834**	**1698**	**329**
Overall, SVI (%)				
	N (%)	**N (%)**	**N (%)**	**N (%)**
1	7183(18.45)	6948(15.16)	331(19.49)	27(8.21)
2	6228(16.00)	7486(16.33)	280(16.49)	46(13.98)
3	5864(15.06)	7372(16.08)	252(14.84)	68(20.67)
4	5836(14.99)	7012(15.30)	294(17.31)	61(18.54)
5	5266(13.53)	7411(16.17)	228(13.43)	76(23.10)
Missing	8558(21.98)	9605(20.96)	313(18.43)	51(15.50)
Total	**38935**	**45834**	**1698**	**329**

Table 3. Details the training and test dataset breakdown by sex assigned at birth, alongside precision and accuracy metrics for the model's classification performance.

igned at birth	**g data (85%)**	**ta (15%)**	**on**	**acy**
Male	2888 (1444 with and without each)	508 (254 with and without each)	Neg: 80% Pos: 80%	80%
Female	560 (280 with and without each)	98 (49 with and without each)	Neg: 80% Pos: 78%	80%

FIGURES

Figure 1. Shows the methodology for matching SendSS and eHARS datasets from Fulton County Georgia from 2010 to 2021. From 132,928 STI cases, 5,729 were excluded. Transposition from case-based to patient-based led to 85,224 individuals, which were then matched to a Social Vulnerability quintile and categorized by sex assigned at birth and documented HIV status.

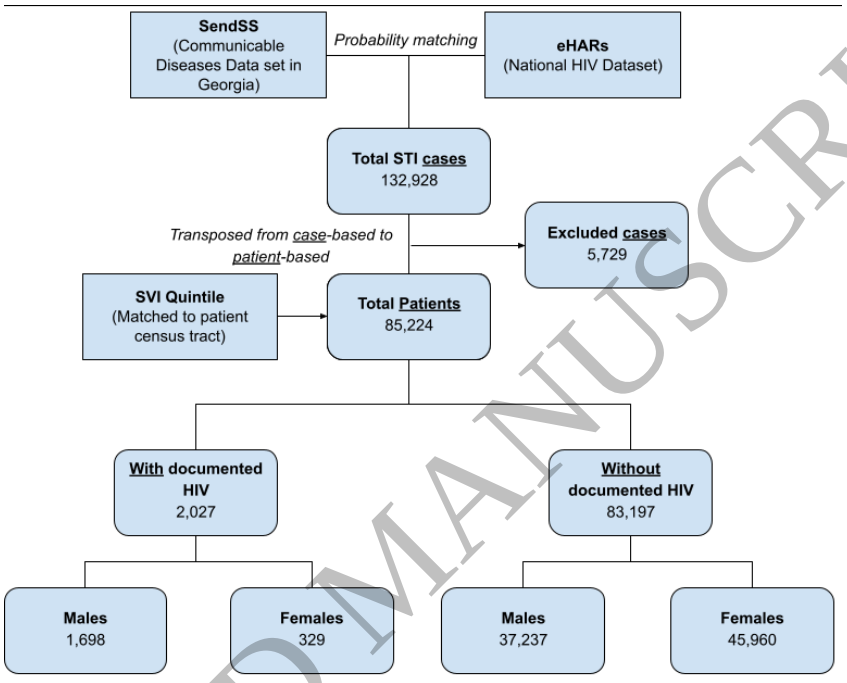


Figure 2. Shows confusion matrices for various machine learning algorithms. True positives, true negatives, false positives, and false negatives are reported for GradientBoostedTrees, Naive Bayes, Logistic Regression, Nearest Neighbors, and Random Forest, with accuracy scores below each matrix. *Similar performance metrics were seen in the Female subgroup (not displayed in this figure).

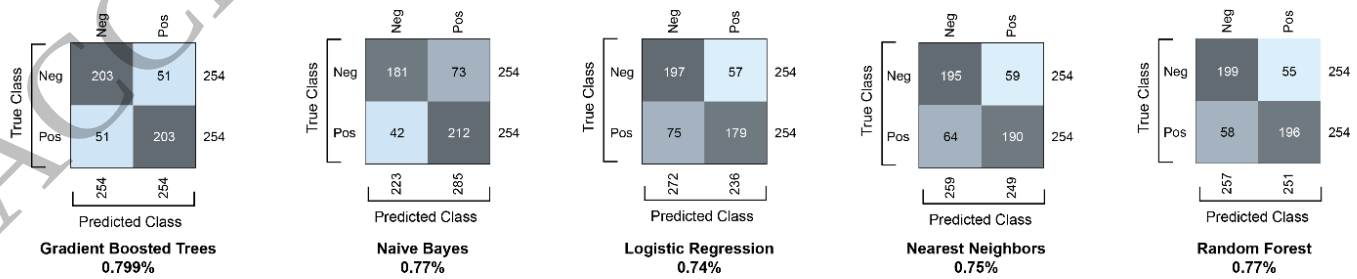
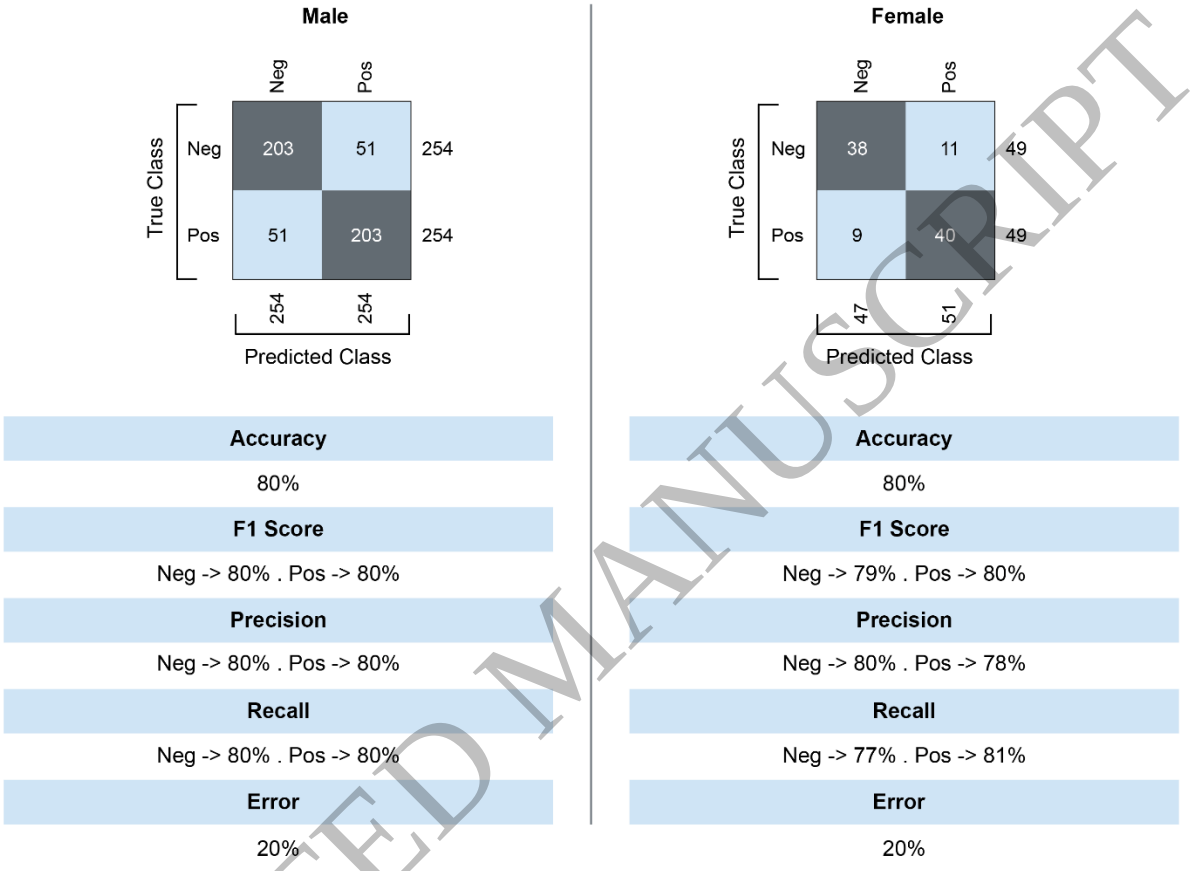


Figure 3. Gradient boosted trees confusion matrices present the performance for males (LEFT) and females (RIGHT) both with an accuracy of 80% with balanced precision and recall across classes. Both models exhibit comparable error rates.



Male	Female
Accuracy	**Accuracy**
80%	80%
F1 Score	**F1 Score**
Neg -> 80% . Pos -> 80%	Neg -> 79% . Pos -> 80%
Precision	**Precision**
Neg -> 80% . Pos -> 80%	Neg -> 80% . Pos -> 78%
Recall	**Recall**
Neg -> 80% . Pos -> 80%	Neg -> 77% . Pos -> 81%
Error	**Error**
20%	20%

Figure 4. Displays the influential features for our predictive model males (left) and females (right). Each bar represents a feature's influence on the model's predictions. Longer bars indicate greater influence.