Machine learning approaches to identify communities with higher HIV prevalence using social economic and behavioral data in resource-limited settings

IDM symposium – Seattle, Washington
October 1st,2024

Masabho Peter Milali, PhD, MSc
Department of Population Health
New York University, Grossman School of Medicine
masabho.milali@nyulangone.org

NYU Langone Health

Introduction

- Measures of HIV prevalence are used to help public health officials, researchers, and policy makers monitor the epidemic, evaluate the impact of interventions, and assist in the identification of sites for HIV prevention trials

- However, current estimation of HIV prevalence strongly depends on the availability of HIV biomarkers, which are challenging and expensive to collect, especially in resource-limited settings

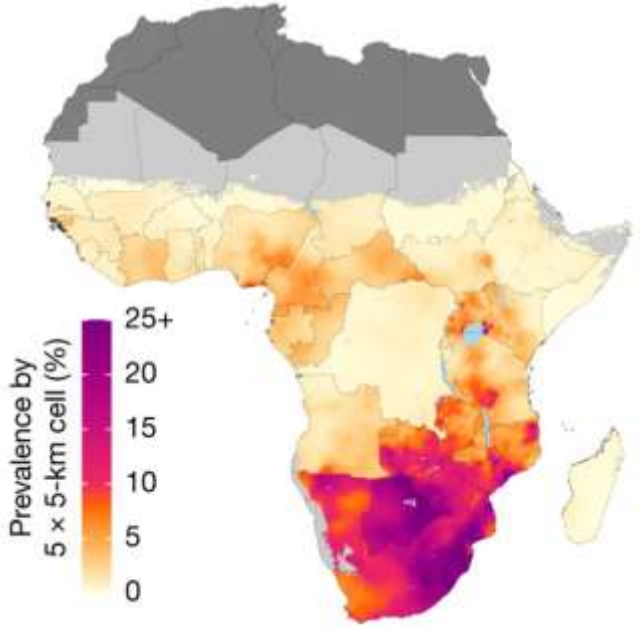


Figure 1. HIV prevalence in sub-Saharan Africa among adults aged 15–49 (Dwyer-Lindgren et al., 2019)

NYU Langone Health

Study objectives

- Can we accurately identify communities with higher HIV prevalence using socio-economic, behavioral, and other community-level data when HIV biomarkers are not available?
  - Partial Least Square (PLS) analysis
  - Decision Tree analysis (Random Forest)

Data

- We utilized data from the Population-based HIV Impact Assessment (PHIA) surveys conducted in Zambia and Kenya
- The PHIA surveys are representative household and individual surveys that provide comprehensive data on socio-economic and behavioral factors related to HIV
- Individual-level data are collected through structured interviews, capturing information on demographics, HIV risk factors, and household characteristics
- After the interviews, consenting individuals provide a blood sample for analysis of HIV-related markers, such as serostatus, CD4 counts, and viral load—forming the HIV biomarker dataset.

Data

- We generated three distinct datasets for model training and testing
- ZAMPHIA: Contains predictor variables from Zambia PHIA surveys.
- KENPHIA: Comprises predictor variables from Kenya PHIA surveys
- ZAKEPHIA: A combined dataset with shared predictors from both surveys, allowing for harmonized model training and testing.

Generating predictor variables from PHIA surveys

- Computed proportions of individuals with feature X per enumeration area (EA), using the total number of qualified individuals as the denominator
- An EA is a geographic area with a 200m radius in urban areas and 1,000m in rural areas, serving as the primary sampling unit
- Examples of questions from surveys that were used to generate variables
  - Have you ever sold sex for money?
  - How many drinks containing alcohol do you take on a typical day?
  - Do you believe women who carry condoms have sex with a lot of men?
  - Are you circumcised?
  - Is your partner living with you now or they stay elsewhere?
  - Does your household receive any form of financial support?

$$\text{E.g., Proportion of circumcised in EA} = \frac{\text{Sum of circumcised males in an EA}}{\text{Total number of males in an EA}}$$

Compute HIV prevalence of enumeration areas

- Using PHIA biomarker data, HIV prevalence was calculated for each enumeration area (EA) by dividing the number of positive for HIV cases by the total individuals who consented to provide blood samples

- Individuals without recorded HIV status were excluded from the calculation

- EAs with prevalence $\geq$ 0.1 were labelled as hotspots (repeated with thresholds 0.05, 0.07, 0.15)

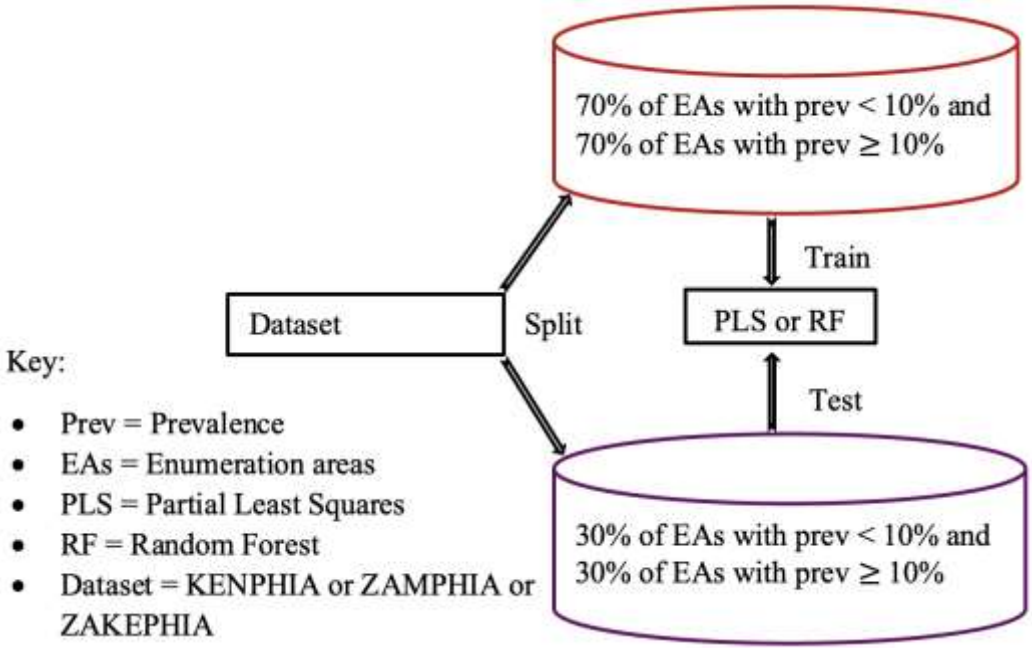
Distance of each EA to health facilities

- We used SciPy library in python to calculate distances from each EA to the nearest health facilities using their geo-coordinates
- Geo-coordinates of each EA were provided by the PHIA surveys, while those for health facilities in Kenya and Zambia were obtained from a published manuscript
- This data allowed us to assess the impact of healthcare access on community-level HIV prevalence

Model training

- EAs in each dataset were categorized based on HIV prevalence – those below 10% were labeled ”coldspots” (0), while those at 10% or higher were labeled “hotspots” (1)

- We merged and randomized the EAs within each dataset into training (70%) and testing (30%) sets, with a total of 511 EAs in ZAMPHIA, 798 in KENPHIA, and 1,309 in ZAKEPHIA



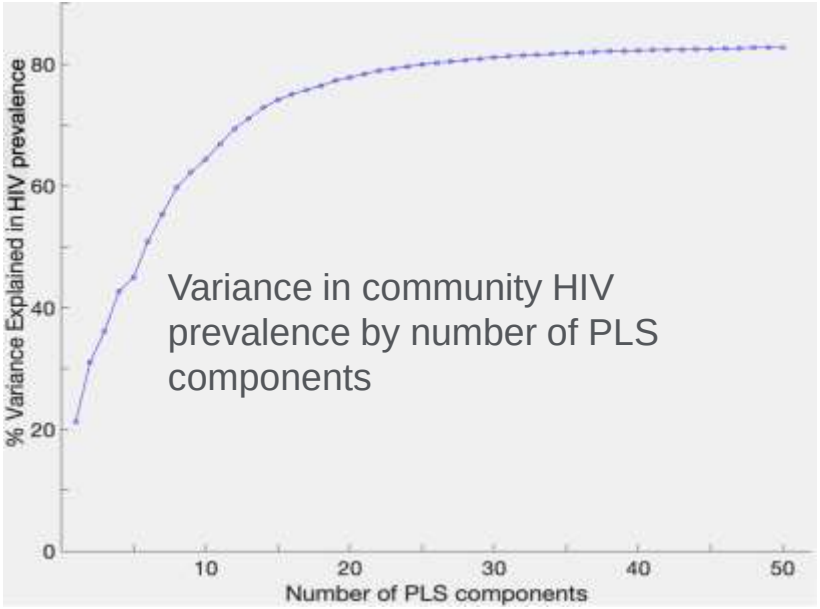
NYU Langone Health

Methods

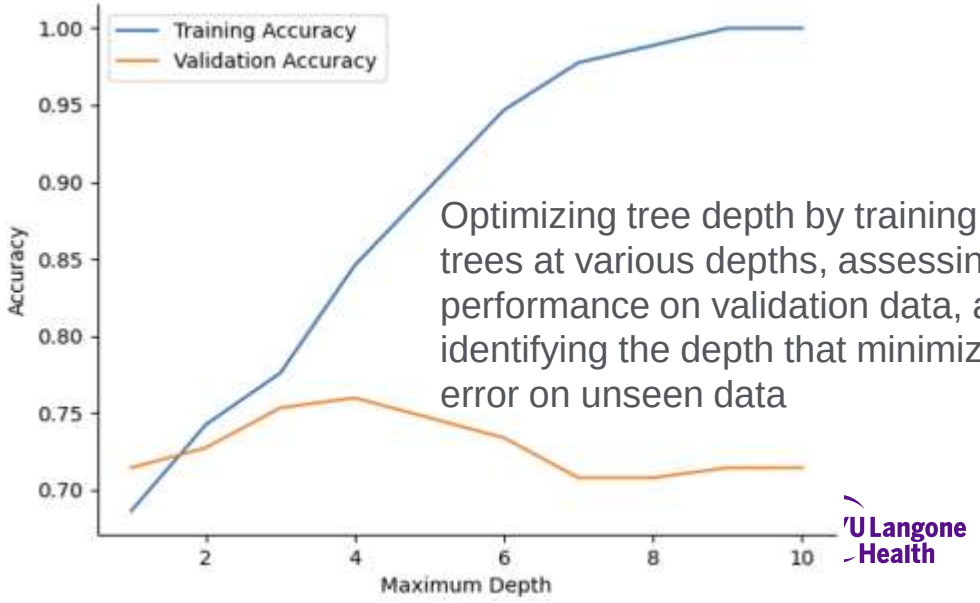
Model training

- Each dataset trained a PLS model with 15 components using ten-fold cross-validation and a RF model with 100 trees and a maximum depth of 4

- The selection of 15 components for the PLS model was based on the variance explained in the response variable relative to the number of components, while the RF model's depth of 4 was determined through cross-validation



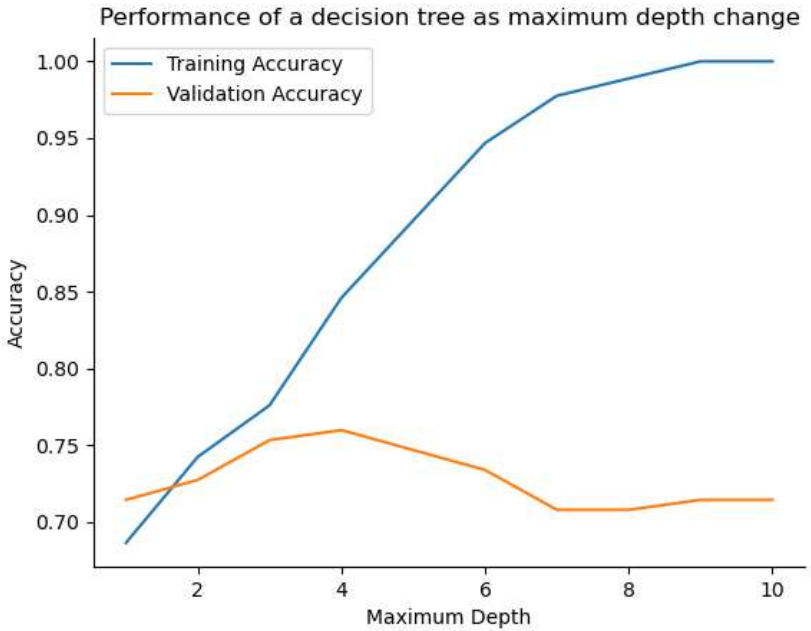
Variance in community HIV prevalence by number of PLS components

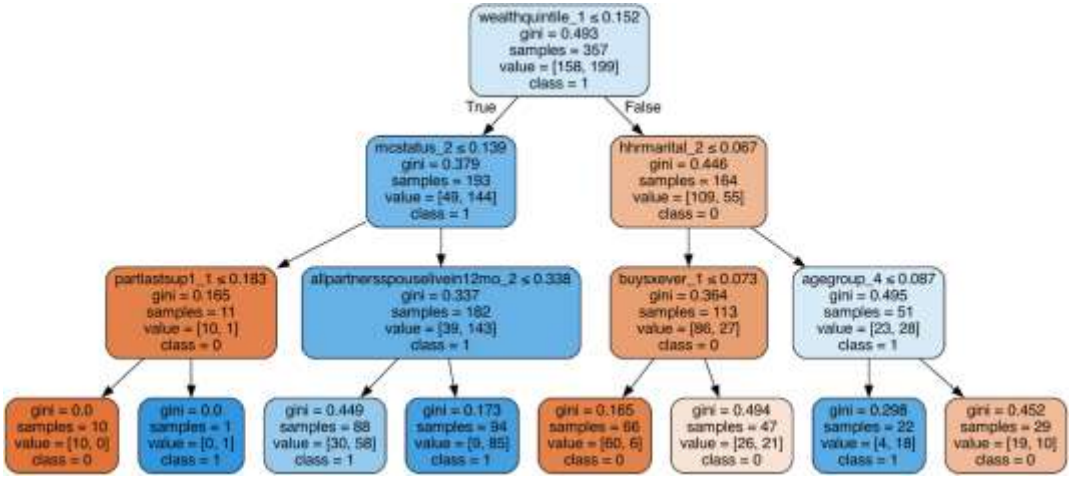


Optimizing tree depth by training trees at various depths, assessing performance on validation data, and identifying the depth that minimizes error on unseen data

NYU Langone Health

Example one decision tree





NYU Langone Health

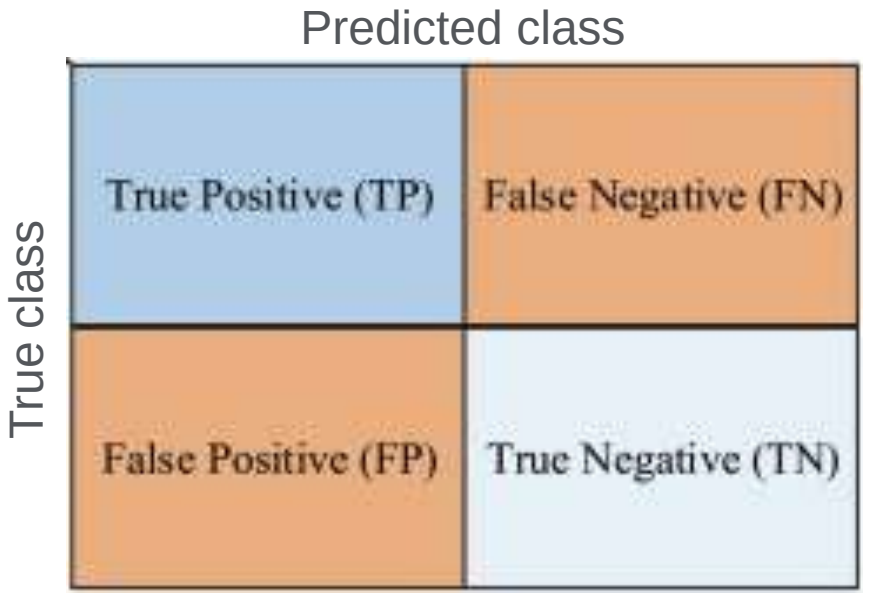
Model training

- To test model robustness on HIV prevalence thresholds, we repeated training and testing using thresholds of 5%, 7%, and 15%, similar to the 10% approach

- To evaluate the extrapolation capabilities of the PLS and RF104 models on datasets with different sample characteristics, we cross-tested the models by applying the Zamphia-trained model to Kenphia and vice versa

Metrics used to score the model performance

1. $\text{Sensitivity} = \frac{\text{True Positives (TP)}}{\text{Actual Positives (TP+FN)}}$
2. $\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{Actual Negatives (TN+FP)}}$
3. $\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+TN+FP}}$, and
4. $\text{Precision} = \frac{\text{T}P}{\text{TP+FP}}$,

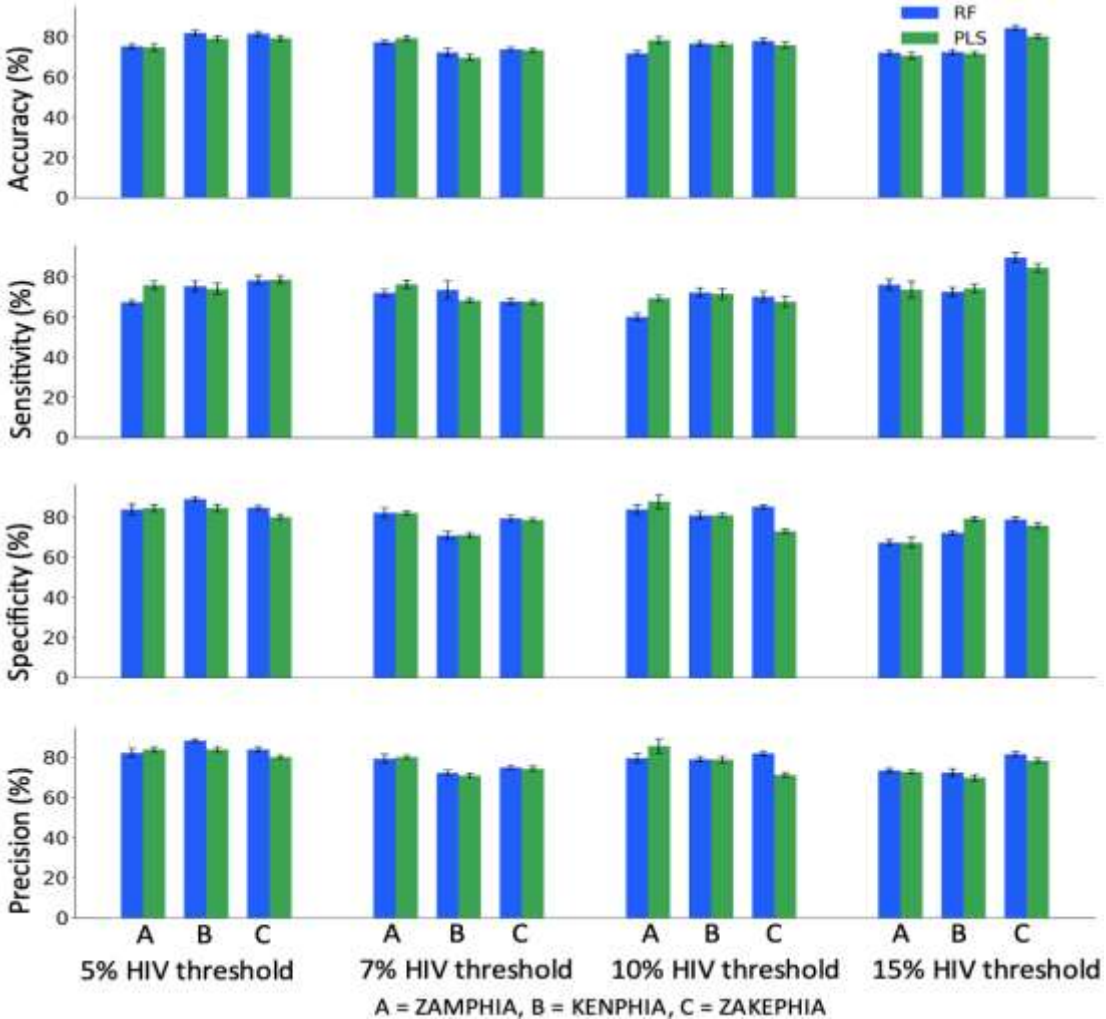


Results

Out of sample testing

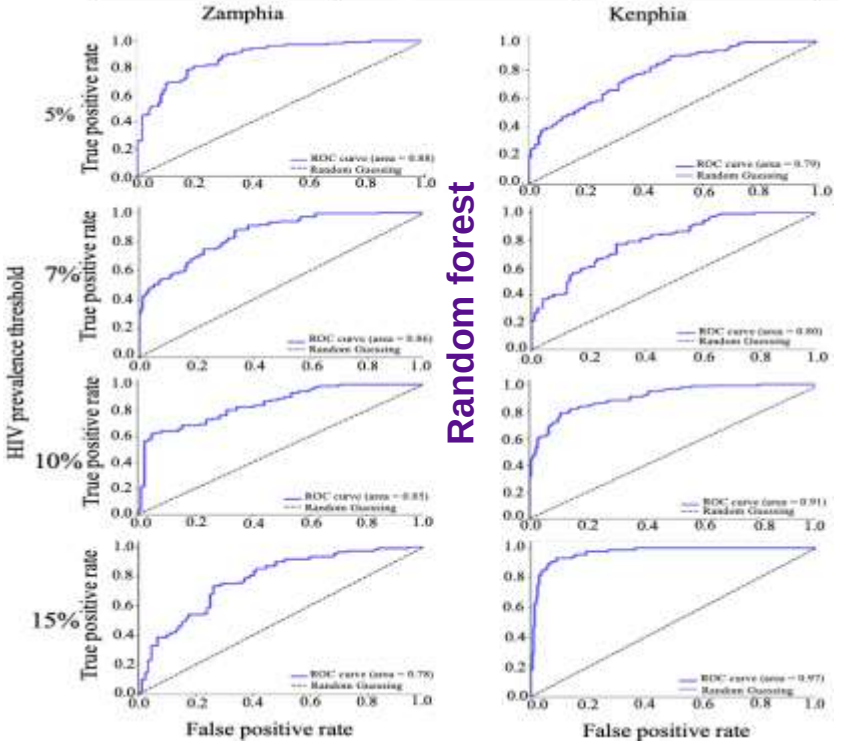
- All models demonstrated robust accuracy (76 ± 5%) across datasets, with variations at different thresholds

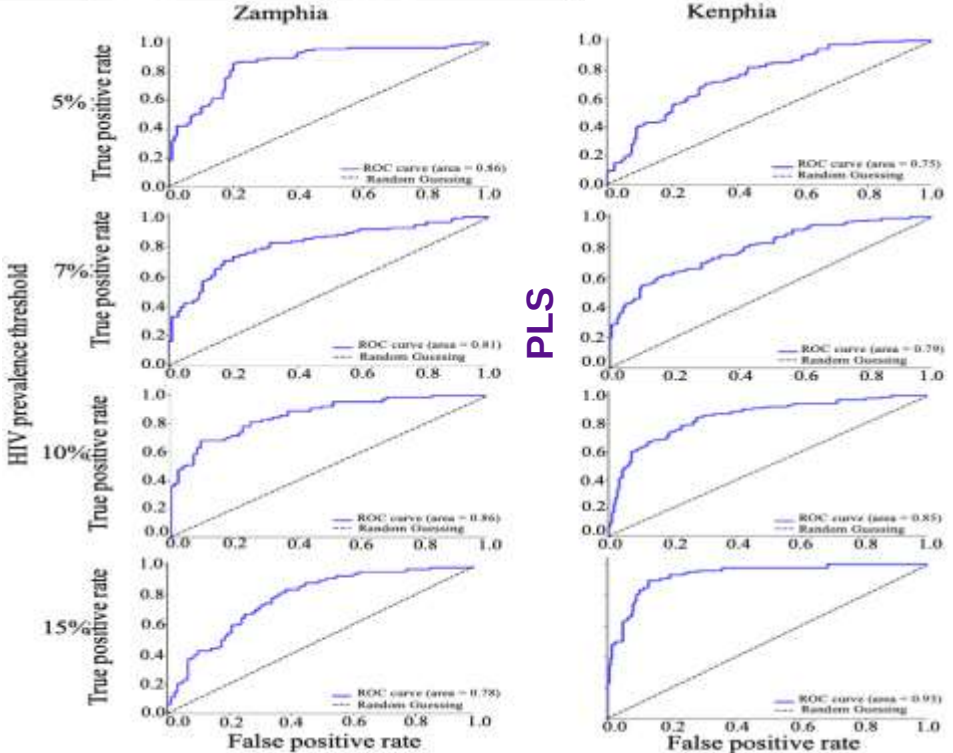
- RF models slightly outperformed PLS models especially at 15% HIV prevalence threshold



Results – ROC curves and AUC

Dataset	Model type	HIV prevalence threshold 5%	7%	10%	15%
ZAMPHIA	RF104	0.88	0.86	0.85	0.78
	PLS	0.86	0.81	0.86	0.78
KENPHIA	RF104	0.79	0.80	0.91	0.97
	PLS	0.75	0.79	0.85	0.93

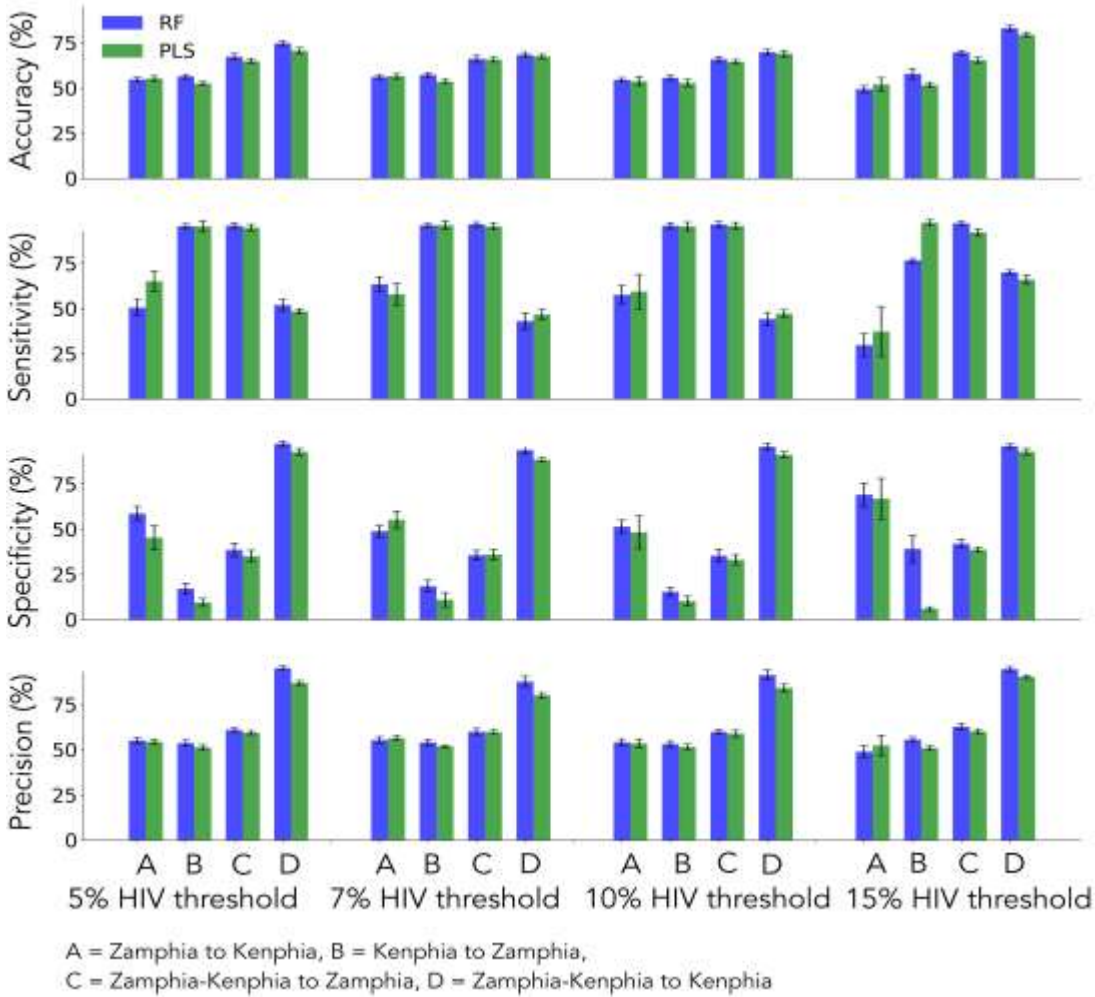




Results

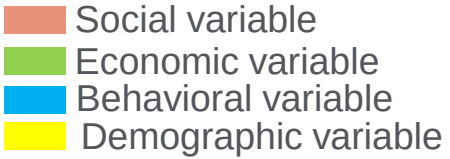
Cross-testing models on datasets from different geographic regions

- The models achieved an overall accuracy of approximately 55%
- Sensitivity for tests A and D, as well as specificity for tests B and C, was below 50%
- Test D consistently outperformed the other tests
- These results are impressive, as the models were tested on datasets that included variables not used during training
- This suggests a low likelihood of overfitting



Results -

Important variables



Kenphia

- Protestants
- Got in relationship for financial support
- Married/cohabiting/living together
- Uncircumcised
- Experienced physical or sexual violence
- Had first intercourse experience before age 15
- 25-29 years of age

Zamphia

- Married at age less than 18 years
- Uncircumcised
- No condom at last sex with partner
- Lowest wealth quintile
- No agriculture land
- Secondary level education
- Partner(s) outside of marriage in past year
- No condom at last sex in past year

NYU Langone Health

Conclusion

- The study shows that socioeconomic and behavioral variables can identify communities with higher HIV prevalence through machine learning (ML)
- This approach underscores the potential of data-driven strategies to inform health policy and advance efforts against infectious and non-communicable diseases across the continent
- Applying ML and AI to health and non-health datasets enables adaptive frameworks that address evolving epidemiological trends and improve health outcomes for vulnerable populations
- Integrating diverse data sources fosters more accurate modeling, providing critical insights for targeted interventions and resource allocation in public health

Acknowledgments

NYU Grossman School of Medicine

Anna Bershteyn (PI)
Hae-Young Kim
Frey Assefa
Shiying You
Daniel Citron
David Kaftan
Ingrida Platais
Neha Kansal
Kasturi Bhamidipati
Sulani Nyimbili
R. Scott Braithwaite

Strathmore University

Samuel Mwalili
Duncan Gathungu

University of Cincinatti

Diego F. Cuadros

Center for Infectious Disease Research in Zambia (CIDRZ)

Izukanji Sikazwe
Sulani Nyimbili

Funding source

National Institutes of Health

NYU Langone Health