PREDICTING CLIENT RETENTION IN AN URBAN HIV CLINIC – A MACHINE LEARNING APPROACH

JONATHAN MELVIN IKAPULE

M23M19/273

A DISSERTATION SUBMITTED TO THE FACULTY OF ENGINEERING, DESIGN AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF A DEGREE OF MASTER OF SCIENCE IN DATA SCIENCE AND ANALYTICS OF UGANDA CHRISTIAN UNIVERSITY

May, 2025

UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

Declaration

I declare that I am the exclusive author of this dissertation and that none of its parts have been previously published or submitted for the award of any degree or qualification.

I confirm that, to the best of my knowledge, my dissertation does not infringe upon any copyrights or violate proprietary rights.

Furthermore, I acknowledge and fully reference any ideas, techniques, quotations, or other content from the work of others, whether published or unpublished, as per standard referencing practices.

Signature: 

Name: Jonathan Melvin Ikapule

Date: May 2025

Approval

This is to certify that this research project titled "Predicting Client Retention in an Urban HIV Clinic - A Machine Learning Approach" has been submitted with my approval as the University supervisor.

Name: Dr Kimbugwe Nasser

Signature: 

Date: May 2025

Abstract

Retention in HIV care is critical to viral suppression, improved health outcomes, and reduced transmission; however, retention rates remain suboptimal in urban Uganda, with some studies reporting rates below 60%. This study aimed to identify retention predictors and develop a machine learning model to predict retention among people living with HIV (PLHIV) using routinely collected patient-level data. A retrospective cohort study was conducted using data from electronic medical records (EMR) from three urban HIV clinics in Kampala (January 2021 - December 2023). Clients who died or were transferred out were excluded, yielding 22,213 clients. Data included demographic, clinical, and visit-related variables, as well as engineered features like duration on antiretroviral therapy, distance to clinic, and viral suppression history. Retention was defined as attending a scheduled appointment within 90 days. Six classification algorithms were trained and evaluated using a 70:30 split and SMOTE (a technique to balance data). Accuracy, precision, recall, and F1 score assessed model performance. XGBoost outperformed other models, achieving an accuracy of 88% and an F1 score of 0.85. Key predictors, identified using SHAP values for feature importance, included duration on ART, weight, age, baseline CD4, distance to the clinic, and ART adherence. These findings demonstrate the feasibility of using EMR data and machine learning to support data-driven decision-making in HIV programs. Machine learning models integrated into EMR systems can enable real-time identification of clients at risk of disengaging from care, guiding targeted interventions. This study highlights the potential of data science to improve HIV service delivery, although further validation in diverse contexts is needed.

Keywords: *Antiretroviral Therapy, Classification, EMR, Retention, SHAP, SMOTE, Supervised Learning, XGBoost, Urban Clinic, Uganda.*

Contents

List of Tables

List of Figures

LIST OF ACRONYMS

AIDS	Acquired Immunodeficiency Syndrome
ART	Antiretroviral Therapy
AUC	Area Under the Curve
CCLAD	Community Client-Led ART Delivery
CDC	Centers for Disease Control and Prevention
CDDP	Community Drug Distribution Point
CSV	Comma-Separated Values
EMR	Electronic Medical Record
HIV	Human Immunodeficiency Virus
KNN	k-Nearest Neighbors
LTFU	Loss to Follow-Up
ML	Machine Learning
MOH	Ministry of Health
PLHIV	People Living with HIV
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	SHapely Additive exPlanations
SMOTE	Synthetic Minority Oversampling Technique
SMS	Short Message Service
SSA	Sub-Saharan Africa
SVM	Support Vector Machine
UNAIDS	The Joint United Nations Programme on HIV/AIDS
XGBoost	eXtreme Gradient Boosting

CHAPTER ONE: INTRODUCTION

1.1 Background

The HIV/AIDS crisis serves as a testament to the profound impact that infectious diseases can have on public health worldwide. The onset of the HIV epidemic in the 1980s heralded one of the most complicated public health challenges of our times. More than 86 million people have contracted HIV since its discovery, and currently, around 39 million people live with the virus. Tragically, millions have lost their lives to AIDS-related complications since the outbreak began UNAIDS2023. Sub-Saharan Africa is still one of the most affected regions, with approximately two-thirds of all PLHIV globally residing in this region, according to Moyo et al. (2023). However, HIV/AIDS is still a public health threat in other world regions, including Asia, Eastern Europe, and Latin America.

Despite the scale of the HIV/AIDS epidemic, remarkable strides have been made in the development of ART and other prevention modalities. HIV infection is no longer the virtual death sentence it was, but is now a chronic, manageable condition for many individuals Broder (2010) since the ART rollout. However, challenges such as treatment access, adherence, and retention in care continue to hinder efforts to achieve epidemic control.

UNAIDS has set 3 goals to achieve HIV epidemic control by 2030. The third goal is to ensure that 95% of all PLHIV who know their status and are on ART achieve viral suppression UNAIDS (2023). Viral suppression, according to the CDC (2023), refers to when a PLHIV has less than 200 copies of HIV per millilitre of blood. To achieve viral suppression, PLHIV need sustained and good adherence to ART; any interruption in this through missed drug pickup appointments and LTFU can have negative effects, including suboptimal viral suppression, leading to increased morbidity and mortality Stricker et al. (2014).

In Uganda, like many other countries heavily impacted by HIV/AIDS, achieving

and maintaining viral suppression among PLHIV is a pressing concern. Urban areas, such as Kampala, face unique challenges related to HIV care delivery, including high patient volumes, limited healthcare resources, and socioeconomic disparities Hardon et al. (2007). Although access to ART has been expanded and HIV care services improved, retention in care rates in urban HIV clinics remain suboptimal Unge et al. (2010), with a worrying proportion of PLHIV experiencing interruptions in treatment, resulting in LTFU for some.

It is a standard procedure for HIV programs to implement tracking mechanisms for patients who miss scheduled clinic appointments to reduce the number eventually becoming LTFU. Follow-up methods include phone calls; home visits, which are used when a phone call is unsuccessful. However, studies indicate dismal performance with less than 25% of clinics implementing timely phone calls and 12.5% conducting home visits Etoori et al. (2020). These results indicate a need for more proactive approaches in identifying clients at higher risk of missing appointments and becoming LTFU.

The adoption of EMR systems has provided healthcare providers with a wealth of patient medical records in formats that can easily be used to attain patient and population-level insights. This, coupled with the proliferation of machine learning predictive algorithms within healthcare Bisaso et al. (2017); Supriya and Deepa (2020), provides an opportunity for more proactive approaches to identifying clients at higher risk of missing appointments and becoming LTFU. Machine learning algorithms, when applied to clinical data, have demonstrated the ability to develop robust risk models that can be instrumental in redefining patient classes Deo (2015) that can then be used to offer targeted interventions. By employing machine learning techniques on retrospective cohort data from three urban HIV clinics in Kampala, this research aimed to predict client retention in a typical HIV program, which will enable healthcare providers to offer targeted interventions to those most at risk.

1.2 Problem Statement

Retention in care is an important health outcome for not only the individual health of people living with HIV but also a public health concern Kumar et al. (2020). Accordingly, retention and accessing care are key to any public health plan to control HIV transmission. In Uganda, retention in care among PLHIV is still a challenge, with Nimwesiga et al. (2023) indicating LTFU stands at 20% for some sub-populations. High dropout rates and suboptimal viral suppression not only compromise individual health outcomes but also pose public health risks through increased potential for transmission. Understanding the complex barriers and facilitators affecting retention is crucial for developing effective interventions. Furthermore, leveraging machine learning to predict these outcomes can enhance proactive patient management and support. However, there is a lack of research utilising machine learning models to predict client retention specifically in urban HIV clinic settings in Uganda. This study aimed to fill this gap by identifying key factors influencing retention and evaluating the performance of machine learning models in predicting these critical outcomes, ultimately improving care strategies and health outcomes for PLHIV in urban areas.

1.3 Main Objective

To develop a machine learning model to predict retention in an urban HIV clinic in Uganda.

1.4 Specific Objectives

The specific objectives of this study were:

1. To determine predictors of retention for PLHIV attending an urban HIV clinic in Uganda.

2. To develop and train a machine learning model that predicts retention for PLHIV attending an urban HIV clinic in Uganda.

3. To validate the performance of a machine learning model that predicts retention for PLHIV attending an urban HIV clinic in Uganda.

1.5 Research Questions

1. What are the key predictors of retention for PLHIV attending an urban HIV clinic in Uganda?

2. How effective is a machine learning model developed to predict retention for PLHIV attending an urban HIV clinic in Uganda?

3. How well does the developed machine learning model perform in predicting retention for PLHIV attending an urban HIV clinic in Uganda?

1.6 Significance of Study

This project aimed to improve HIV care for people living with HIV (PLHIV) in urban areas by identifying what factors influence their retention in care and treatment. Regularly clinic attendance usually results in sustained ART which in turn leads to viral suppression, one of the preferred outcomes for PLHIV. We developed and tested machine learning models to predict which patients were most likely to be retained in care. This is expected to help healthcare providers offer timely and better individual-level support to those in most need especially in resource limited settings like Uganda. The project findings were expected to improve health outcomes, reduce HIV transmission, and provide insights for better public health strategies.

1.7 Scope of Study

The study was conducted at St Balikuddembe Market Clinic, Dr Charles Farthing Memorial Clinic and China-Uganda Friendship Hospital-Naguru, three urban HIV clinics located in Kampala, Uganda. It used routinely collected clinical and demographic data for clients who attended any of the clinics between January 2021 and December 2023. The study fell within the public health and data science disciplines and focused on the development of machine learning models to predict retention in care. The study did not explore qualitative factors that could influence retention in care.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

Retention in HIV care is essential for the success of ART in PLHIV attending urban clinics, where high patient mobility, clinic overcrowding, and socio-economic barriers contribute to dropout rates as high as 40% within two years of ART initiation Koole et al. (2014); Fox and Rosen (2015). Poor retention undermines treatment outcomes, increases risks of drug resistance, and fuels HIV transmission. Supervised ML classification algorithms, such as Random Forests (RF) and XGBoost, offer a promising approach to predict retention by modelling variables like age, distance to clinic, CD4 count, and viral load (VL) suppression status. However, their application to urban HIV clinics remains underexplored, with few studies leveraging comprehensive patient data to identify at-risk individuals. This chapter reviews the literature in three key areas: (1) retention challenges in urban HIV care; (2) supervised ML applications in healthcare; and (3) studies using ML to predict HIV retention, identifying gaps addressed by this study's use of multiple classification algorithms and urban-specific variables.

2.2 Retention in HIV Care

2.2.1 Definition and Metrics of Retention

Retention in care is a key component for successful HIV management, particularly in SSA, where the burden of HIV is highest. Retention in care refers to the continued engagement of PLHIV in clinical services, involving regular clinic visits, viral load monitoring, and adherence to ART. Retention in HIV is essential for achieving viral suppression, reducing morbidity, and preventing transmission Organisation (2012). Common metrics include loss to follow-up (LTFU), defined as missing clinic visits for 90 days or more, visit consistency, and the proportion of scheduled visits attended Stricker et al.. In urban settings, retention is often

measured by adherence to quarterly visits due to high patient volumes and mobility Camlin and Charlebois (2019). These metrics provide a basis for assessing retention but are limited by retrospective data, highlighting the need for predictive models to identify at-risk patients proactively.

2.2.2 Factors Influencing Retention

Retention in urban HIV clinics is challenged by a complex interplay of clinical, social, and logistical factors, with retention rates being around 71% two years post-ART initiation Fox and Rosen (2015). Sebunya et al. (2013) found that in Ugandan urban clinics, clients newly initiated on ART were more likely to disengage within the first year, particularly those with low CD4 counts or poor ART adherence. Distance to clinic significantly impacts retention, with Terzian et al. (2018) reporting a 30% higher dropout risk for urban patients travelling over 8 km in a Washington, DC, cohort. Gender and socio-economic factors also play a role; Rachlis et al. (2017) noted that urban men with unstable employment had a 25% higher dropout risk compared to women. Stigma and limited treatment education are known to increase levels of missed appointments, particularly in urban settings where social and economic pressures are amplified Turan et al. (2017). History of regimen change and VL non-suppression, which could point to poor treatment adherence, are additional risk factors, with studies showing that patients with prior VL non-suppression are more likely to drop out Yotebieng et al. (2019). Tuberculosis Preventive Therapy (TPT) status and duration on ART also influence outcomes, as longer ART duration correlates with better retention Rangaka et al. (2014). Strategies like SMS reminders, peer support, and differentiated service delivery (DSD) models have improved retention by up to 24% in urban clinics Steward et al. (2021); Uetela et al. (2023). However, traditional methods are poor at proactively identifying patient categories at increased risk for early intervention and only come into play once an event has occurred. Supervised ML can enhance the identification of patient categories at higher risk in urban settings by

using features like age, distance, CD4 count, ART adherence, and VL suppression status.

2.2.3 Consequences of Poor Retention

Poor retention in HIV care has severe consequences for both individual and public health outcomes. Poor retention is associated with viral load non-suppression, increased drug resistance, and higher transmission risks, with Nanyeenya et al. (2023) estimating that patients with a history of VL non-suppression double the risk of onward HIV transmission in urban communities (aHR = 2.1, 95% CI: 1.6–2.8). Low CD4 counts and poor ART adherence usually lead to increased morbidity risks, leading to higher hospitalisation rates and increased healthcare costs Owachi et al. (2024). In urban clinics, where higher patient volumes result in straining of resources, poor retention also increases clinic inefficiencies, diverting resources from stable patients Owachi et al. (2024). These consequences highlight the urgent need for predictive tools to identify patients at higher risk of disengaging from care to implement measures to retain them, particularly in urban HIV settings.

2.3 Machine Learning in Healthcare

2.3.1 Overview of ML in Healthcare

Machine learning involves the use of algorithms that learn patterns from data to make predictions, with supervised classification algorithms like Logistic Regression (LR), Random Forests (RF), XGBoost, Support Vector Machines (SVM), K-Nearest Neighbours (KNN), and Naïve Bayes widely used in healthcare. These algorithms identify patterns in complex relationships in patient data to predict outcomes such as hospital readmissions or disease progression. For example, RF achieved 85% accuracy in predicting hospital readmissions in urban settings Adhiya et al. (2024) XGBoost improved diabetes management predictions with 90%

AUC Wang et al. (2020). LR and Naïve Bayes are valued for interpretability, while SVM and KNN excel in handling non-linear data. These algorithms leverage features like age, clinical markers, and social factors, making them suitable for predicting retention in HIV care.

2.3.2 ML for Patient Retention and Engagement

ML has been applied to predict patient behaviours, such as appointment attendance and treatment adherence, in chronic disease management, though HIV-specific studies are limited. Shour et al. (2023) used XGBoost to predict clients likely to miss their appointments, achieving an AUC of 0.83 with features like age, distance to clinic, and history of missing appointments. Similarly, Gichuhi et al. (2023) applied ML to predict tuberculosis treatment adherence in urban settings, using clinical markers like weight and treatment duration, with SVM achieving 91% accuracy. In HIV care, Gwadu et al. (2023) used LR to model ART adherence based on CD4 count and VL history, achieving 75% precision, but lacked urban-specific focus. These studies demonstrate the potential of supervised ML to predict retention-related outcomes using variables like those in this study (e.g., age, distance, CD4, VL status). However, their application to urban HIV clinics is limited, with most studies focusing on general healthcare settings or other diseases, but not HIV.

2.3.3 Limitations of Current ML Approaches

Despite ML's potential, limitations exist in its application to retention in healthcare. Many models suffer from biased or incomplete datasets, particularly in urban HIV settings where data on variables like TPT status or regimen change are often missing . Generalizability is another challenge, as models trained on non-urban or non-HIV populations may not apply to urban HIV clinics with high mobility or socio-economic diversity . Few studies compare multiple algorithms like LR, RF, and XGBoost, limiting insights into optimal approaches for retention prediction.

Additionally, variables like VL history or TPT status are rarely incorporated, reducing model relevance to HIV care . These limitations highlight the need for tailored ML models in urban HIV contexts.

2.4 Gaps in the Literature

The literature reveals significant gaps in both HIV retention studies and ML applications. Most retention studies rely on retrospective analyses, lacking predictive tools to identify at-risk patients in urban clinics Ramachandran et al. (2020). While factors like distance, CD4 count, and VL non-suppression are well-documented, their integration into predictive models is limited, particularly in urban settings with unique challenges like mobility and overcrowding Bond et al. (2018). In ML, studies like Schmalzle et al. (2024) demonstrate the efficacy of XGBoost and RF in urban healthcare, but their application to HIV retention is limited, with most studies focusing on diabetes or tuberculosis Oliwa et al. (2021). No studies in Uganda have comprehensively compared supervised algorithms or incorporated specific variables like TPT status, regimen change, or VL history to predict retention in HIV care. This study addressed these gaps by applying multiple classification algorithms to a rich dataset of specific variables, including age, distance, CD4 count, and VL suppression status, to predict retention in an urban HIV clinic.

2.5 Conclusion

Retention in care for PLHIV attending urban HIV clinics is critical to the success of ART. However, it faces barriers such as distance, socioeconomic factors, and clinical factors such as CD4 count and VL non-suppression. Although supervised ML algorithms have shown promise in predicting healthcare outcomes, their application to urban HIV retention remains limited, with few studies that leverage comprehensive variables or compare multiple algorithms. The identified gaps -

lack of predictive models for urban HIV retention and underuse of variables such as TPT status or regimen change - justify this study's approach. Chapter 3 details how this study applied supervised classification algorithms to selected variables to address the need for tailored retention prediction in urban HIV clinics.

CHAPTER THREE: METHODOLOGY

3.0 Introduction

This chapter details the methods that were used in this retrospective cohort study. It includes the research design, study design, data collection and data analysis procedures. Through a systematic examination of medical records, the study sought to identify and extract key features. ML algorithms were applied to build a model for predicting retention in care, and SHAP (SHapley Additive exPlanations) values were used to identify the key predictors of retention, directly addressing specific objective 1.

3.1 Research Design

Research design served as a guide for how the collection and analysis of data were conducted, with a focus on consistency, reliability and validity throughout the study Jansen (2023). A correlational research design was employed, whereby routinely collected, anonymised, longitudinal data were used to examine associations between predictor variables and the outcome of retention in care. This design was appropriate given the study did not involve any intervention or manipulation of variables, but rather sought to identify naturally occurring relationships within existing clinical data.

3.2 Study Design

This study employed a retrospective predictive design, utilising routinely collected clinical data from 3 urban HIV clinics in Kampala, Uganda. The objective was to identify key predictors of patient retention and develop a machine learning model capable of predicting the likelihood of a PLHIV being retained in care. Through analysing historical patient-level data, the study aimed to obtain insights that can support proactive retention strategies in HIV care and treatment programs.

3.3 Study Setting

The study relied on patient data from three HIV clinics located in Kampala, Uganda, all of which were situated in an urban environment. These clinics included:

- China-Uganda Friendship Hospital (CUFH) – Naguru, a government-run facility
- St. Balikuddembe Market Clinic and Dr. Charles Farthing Memorial Clinic, both of which are NGO-supported facilities.

These clinics provide comprehensive HIV prevention, care and treatment services including HIV testing, ART, adherence counselling, and patient monitoring. As urban clinics, they serve a diverse population of PLHIV, including mobile and hard-to-reach groups. These facilities routinely collect and manage patient-level data through electronic medical record systems, which formed the basis for this study.

3.4 Study Population

The study included all patients who received HIV care services at any of the three selected clinics between January 2021 and December 2023. This included patients who had at least one clinic visit within the specified period.

Clients identified as inactive due to death or formal transfer out to another facility at the time of data extraction were excluded from the dataset. These cases were removed to ensure our outcome variable – retention in care- reflected a patient's actual engagement status rather than outcomes unrelated to clinic performance or patient behaviour.

After applying the exclusion criteria and performing data cleaning, 22,213 clients were included in the final analytical dataset used for model development and evaluation.

3.5 Data Sources and Collection

Data for this study were extracted from UgandaEMR version 4.0.3, an electronic medical records system used by all three participating clinics. UgandaEMR is a customised implementation of OpenMRS supported by Uganda's Ministry of Health for managing patient-level data across HIV programs.

A cohort of patients meeting the inclusion criteria (i.e. clinic visits between January 2021 and December 2023) was generated within the EMR system. Data from this cohort was then exported in MS Excel format for analysis. The following variables were extracted:

- Demographic Information: Clinic ID, Gender, Age, Patient Sub-County
- Clinical and Treatment Information: ART start date, Baseline CD4 count, Previous and Current VL results, Regimen Change date, ART Adherence, Weight, TPT Status, TB treatment status.
- Service Delivery Model: DSDM category
- Support Structures: Patient Contact Information, Treatment Supporter
- Visit and Outcome Data: Visit Dates, Last Appointment Date, Transfer Out Date, Death Date

This comprehensive dataset formed the basis for both descriptive analysis and the development of a predictive model to assess patient retention, which is a well-established clinical outcome within the field Sigaloff and de Wit (2015); Stricker et al. (2014).

3.6 Variables

The primary outcome variable for this study was retention-in-care, operationalised as a binary variable. A patient was classified as LTFU if they had missed their most recent scheduled appointment by 90 days or more at the time of data extraction,

and any follow-up attempts were unsuccessful. Patients who were not LTFU were considered retained in care.

3.6.1 Predictor Variables

A range of demographic, clinical, service delivery, and behavioural variables were used as predictors. These included:

- Demographic: Age, Gender, Sub-County
- Clinical: Baseline CD4 count, ART adherence, Weight, TPT status
- Support Structures: Treatment Support and Patient Contact Information

3.6.2 Feature Engineering

To enhance the performance of the model, several variables were created:

- Duration on ART: Time (in months) between ART initiation and the last recorded clinic visit
- Distance to Clinic: Estimated based on the patient's sub-county of residence and categorized into $< 5km$, 5-10km, 10-20km, 20-30km, and $> 30km$.
- Waiting Time: Inferred from the patient's DSDM category to estimate the time spent during a clinical visit
- Peer Support: Derived from the patient belonging to a peer support group or not
- History of VL Non-Suppression: A binary variable indicating whether the patient ever had a VL $>$ 1000 copies/mL
- Current VL Suppression Status: A binary indicator of whether the most recent VL result was suppressed ($\leq$ 1000 copies/mL)

These engineered variables were included to capture patient behaviour, treatment response, and service delivery factors likely to influence retention outcomes.

3.7 Data Processing and Cleaning

Data preprocessing and cleaning were conducted before model development to ensure quality and consistency. Records with missing key identifiers or invalid values were excluded during the initial cleaning phase. For the remaining dataset, the following procedures were applied:

3.7.1 Missing Data Handling

Variables with more than (30%) missing values were excluded from the analysis. For the remaining variables, missing values were imputed using the mode of the respective variable, as missing values were considered to be at random and the proportion was low, making simple imputation appropriate.

3.7.2 Encoding

Ordinal and binary variables were encoded using label encoding to preserve the natural order. This applied to variables like ART adherence and Distance to Clinic.

Nominal variables were transformed using one-hot encoding to ensure appropriate representation in machine learning algorithms.

These preprocessing steps were implemented using Python 3 McKinney (2017), with the Pandas and Scikit-learn libraries. The cleaned and transformed dataset was then used for training and evaluating machine learning models.

3.8 Model Development

Supervised learning classification algorithms (LR, RF, XGBoost, SVM, KNN, Naïve Bayes) were used to develop models before performing model selection Maskew et al. (2022). The data was randomly split into training (70%) and testing (30%) subsets Ombui (2023); Kumar et al. (2020). Given the class imbalance between retained and LTFU patients, SMOTE was applied to the training

set. This ensured that the models were not biased toward the majority class and could better learn the patterns associated with LTFU.

All model training, oversampling, and preprocessing steps were conducted using Python, primarily with the Scikit-learn, XGBoost, and Imbalanced-learn libraries. No manual feature selection was applied in advance; instead, each model was trained using the full set of input variables, including both original and engineered features.

3.8.1 Feature Importance Analysis

SHAP values were computed for the best-performing model following model selection. SHAP values provide a theoretically grounded, model-agnostic framework for interpreting machine learning predictions by quantifying the marginal contribution of each feature to individual predictions, based on cooperative game theory. This approach was preferred over simpler methods such as built-in tree importance scores, as SHAP values account for feature interactions and provide consistent, locally faithful explanations.

For each predictor variable, mean absolute SHAP values were calculated across all test set observations to produce a global feature importance ranking. Variables with the highest mean absolute SHAP values were identified as the strongest predictors of retention in care. Direction of influence (i.e. whether higher values of a predictor increased or decreased the likelihood of retention) was additionally interpreted from SHAP summary plots, providing clinically meaningful insight into each predictor's effect on the outcome.

3.9 Model Evaluation

To assess and compare the performance of each machine learning model, four complementary evaluation metrics were calculated using the test dataset.

Accuracy: The proportion of total correct predictions (both retained and LTFU)

out of all predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where:

- TP: True Positives
- TN: True Negatives
- FP: False Positives
- FN: False Negatives

Precision: The proportion of correctly predicted LTFU cases out of all cases predicted as LTFU.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity): The proportion of actual LTFU cases that were correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1 Score: The harmonic mean of precision and recall, balancing the trade-off between false positives and false negatives.

$$F1 - Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Each of the six models was evaluated using these metrics. Among them, the F1 Score was selected as the primary evaluation metric to identify the best-performing model for predicting retention in care.

All model evaluation and comparison were conducted using Python's Scikit-learn library, with additional visualisation libraries used to support performance analysis in the results chapter.

3.10 Limitations of the Study

The study provided valuable insights into the key predictors of retention in care among HIV patients; several limitations should be acknowledged:

3.10.1 Retrospective Design

Retrospective studies are limited by the quality and completeness of the available data. Missing data points, inaccuracies, and inconsistencies in the EMRs may have introduced bias into the study's findings. Despite efforts to clean and impute missing values, the possibility of residual data quality issues may remain.

3.10.2 Data Imbalance

The study relied on SMOTE for handling class imbalance. While this technique helps mitigate the effects of class imbalance, it cannot eliminate the potential of overfitting.

3.10.3 Assumptions in Feature Engineering

Assumptions were made during feature engineering, particularly in categorising distance from the clinic and estimating waiting times based on DSDM categories, which may not have captured all nuances of patient behaviour or clinic service delivery, which could influence retention patterns.

3.10.4 Exclusion of Certain Patient Populations

Patients who were inactive due to death or transfer out at the time of data extraction were excluded from the study. This exclusion may have overlooked certain patterns of retention or LTFU among these patients, potentially limiting the study's scope.

3.10.5 Model Performance

Although various ML models were tested, there is no guarantee that the best-performing model based on accuracy will necessarily generalise well to other datasets or settings. Further validation with external datasets is recommended for more robust conclusions.

3.11 Ethical Considerations

This study was conducted with adherence to institutional ethical guidelines. Ethical approval was obtained from the relevant ethics committee.

Since it was a retrospective study using secondary data, patient consent was not required. The data used was de-identified to maintain patient privacy and confidentiality. All PII was removed before data extraction, ensuring no individual patient could be identifiable from the dataset.

Additionally, all data was handled securely, with access limited to authorised personnel only. The study was fully in compliance with Uganda's data protection and privacy laws and the ethical standards of Uganda's MOH.

CHAPTER FOUR: RESULTS

4.0 Introduction

This chapter presents the results of the study across three objectives: identifying key predictors of retention, developing and training a machine learning model to predict retention, and validating model performance. A total of 22,213 clients drawn from three urban HIV clinics in Kampala were included in the analysis. Results are organised into four sections: descriptive statistics, data distribution and model performance, key predictors of retention, and model evaluation and interpretation.

4.1 Descriptive Statistics

A total of 22,213 clients were included in the study cohort, drawn from three urban HIV clinics in Kampala: CUFH-Naguru (government-run), St Balikuddembe Market Clinic, and Dr Charles Farthing Memorial Clinic (both NGO-supported). The cohort included all clients who had at least one clinic visit between January 2021 and December 2023, excluding those who were inactive due to death or transfer out at the time of data extraction.

Demographic Characteristics

Table 1 presents a summary of the demographic and clinical characteristics of the 22.312 clients included in the study cohort.

The gender distribution showed a predominance of women (64.6%), which aligns with national trends in HIV care programs, as shown in Figure 1. The client's ages ranged from young adolescents to older adults, with an average age of 40 years, as illustrated in Figure 2. The majority of clients (96%) of the clients had contact information recorded in the EMR system.

	Gender	Age	Patient_Contact	RxSupport	PeerSupport	Approach	Waitingtime	DistancetoClinic	ARTduration	Baseline_CD4	NonSuppHist	VLSuppCurr	RxChangeEver	ARTAdherence	Weight (kg)	TPT_status	HistoryofOIs	HistoryofTransfers	Outcome
count	22,213	22,213	22,213	22,213	22,213	21,984	21,984	22,213	22,213	22,213	21,773	21,587	22,213	22,087	21,846	21,955	22,213	22,213	22,213
unique	2	NaN	2	2	2	2	2	5	NaN	NaN	2	2	2	3	NaN	5	2	2	2
top	Female	NaN	Yes	Yes	No	Facility	Short	5-10km	NaN	NaN	No	Yes	No	Good	NaN	Completed	No	No	Retained
freq	14,358	NaN	21,432	18,636	18,654	18,792	11,967	7,718	NaN	NaN	20,546	20,778	13,468	21,494	NaN	20,382	20,937	19,390	19,695
mean	NaN	41	NaN	NaN	NaN	NaN	NaN	NaN	103	336	NaN	NaN	NaN	NaN	68	NaN	NaN	NaN	NaN
std	NaN	11	NaN	NaN	NaN	NaN	NaN	NaN	52	284	NaN	NaN	NaN	NaN	16	NaN	NaN	NaN	NaN
min	NaN	2	NaN	NaN	NaN	NaN	NaN	NaN	0	0	NaN	NaN	NaN	NaN	5	NaN	NaN	NaN	NaN
25%	NaN	33	NaN	NaN	NaN	NaN	NaN	NaN	65	114	NaN	NaN	NaN	NaN	57	NaN	NaN	NaN	NaN
50%	NaN	40	NaN	NaN	NaN	NaN	NaN	NaN	103	293	NaN	NaN	NaN	NaN	66	NaN	NaN	NaN	NaN
75%	NaN	48	NaN	NaN	NaN	NaN	NaN	NaN	137	477	NaN	NaN	NaN	NaN	76	NaN	NaN	NaN	NaN
max	NaN	91	NaN	NaN	NaN	NaN	NaN	NaN	436	1500	NaN	NaN	NaN	NaN	179	NaN	NaN	NaN	NaN

Table 1: Summary of Patient Characteristics

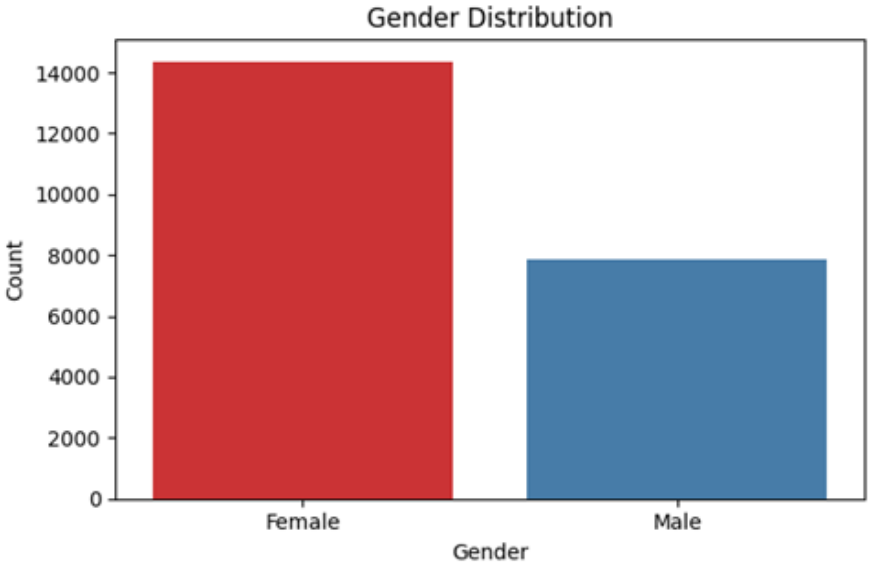

Figure 1: Distribution of Study Population by Gender

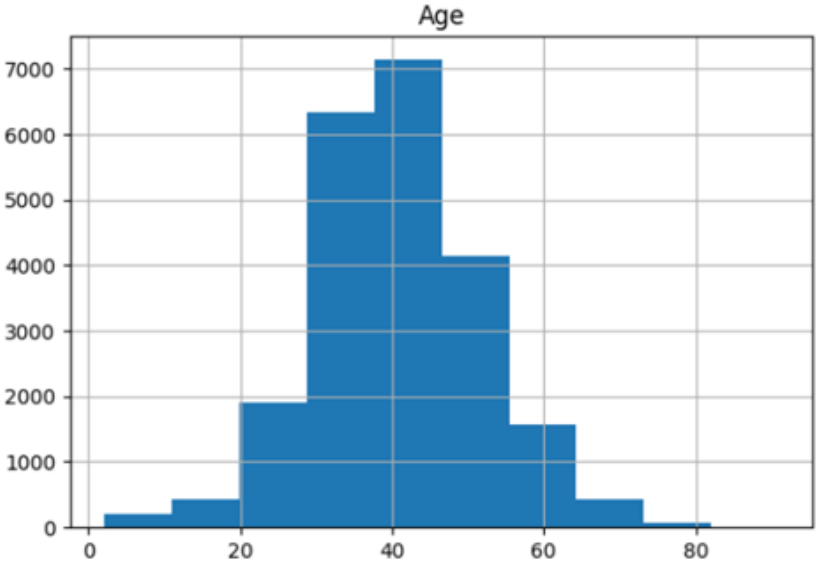

Figure 2: Distribution of Study Population by Age

4.1.2 Clinical Characteristics

Key clinical variables included duration on ART, baseline CD4 count (average of 335 cells), history of opportunistic infections (6%), history of ART regimen changes (39%), history of VL non-suppression (5.6%), current VL suppression status, and ART adherence. Most patients had initiated ART more than a year before the end of the study period, with an average of 8 years on ART, with a substantial proportion of clients, 61%, with no history of regimen change.

Clients were also categorised by their DSDM approach, which included community-based and facility-based models; 84% of clients were in the facility-based models. Additionally, TPT status, weight and treatment supporter presence were also recorded.

4.1.3 Location and Access-Related Variables

The distance to the clinic was estimated using the client's sub-county of residence and categorized into five groups (<5km, 5-10km, 10-20km, 20-30km, and >30km). This proxy variable was used to examine geographical accessibility and its impact on retention. The distribution of clients across distance categories is presented in Figure 3

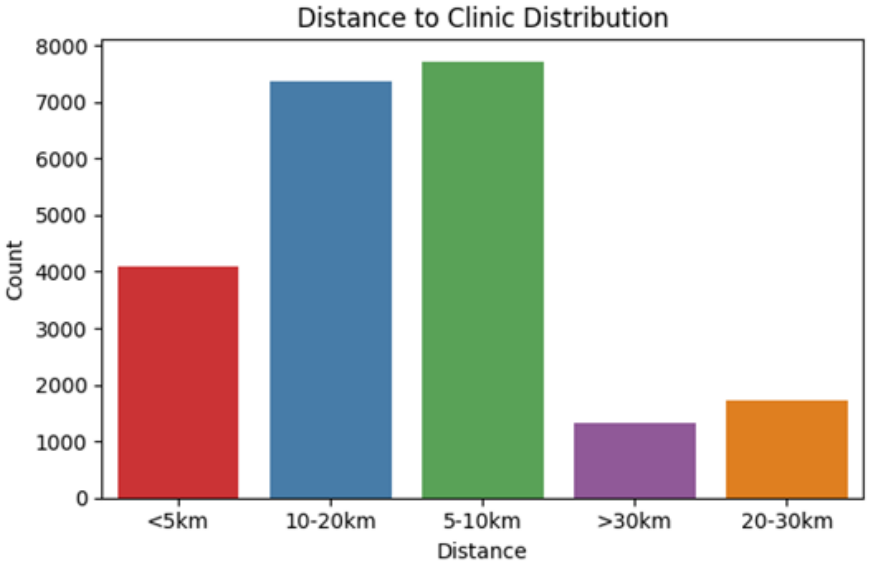

Figure 3: Distribution of Study Population by Distance to Clinic

4.1.4 Summary of Missing Data

Missing data was observed for a small number of categorical variables (less than (3%) across all variables. These were treated by imputing the mode of the respective variable as missing values were considered low and at random. Numerical variables were handled as part of data preprocessing and feature engineering steps described in Chapter 3.

4.2 Data Distribution and Model Performance

4.2.1 Class Distribution

The outcome variable was binary: retained versus LTFU. Of the 22,213 clients analysed, a significantly higher proportion, 89%, were retained compared to those LTFU, indicating a class imbalance, as illustrated in Figure 4. This imbalance was addressed during model training by applying SMOTE exclusively to the training set to generate additional synthetic LTFU cases, while the original class distribution was preserved in the test set to ensure realistic performance estimation.

4.2.2 Model Training and Testing

Six classification algorithms were trained and evaluated on the dataset: Logistic Regression, Random Forest, XGBoost, Support Vector Machines, K-Nearest Neighbors, and Naïve Bayes.

The dataset was split into a 70%-30% training and testing set, respectively. All models were trained using the same training data with SMOTE applied and tested on the untouched 30% test set.

4.2.3 Evaluation Metrics

The models were evaluated using the following metrics: Accuracy, Precision, Recall, and F1 Score.

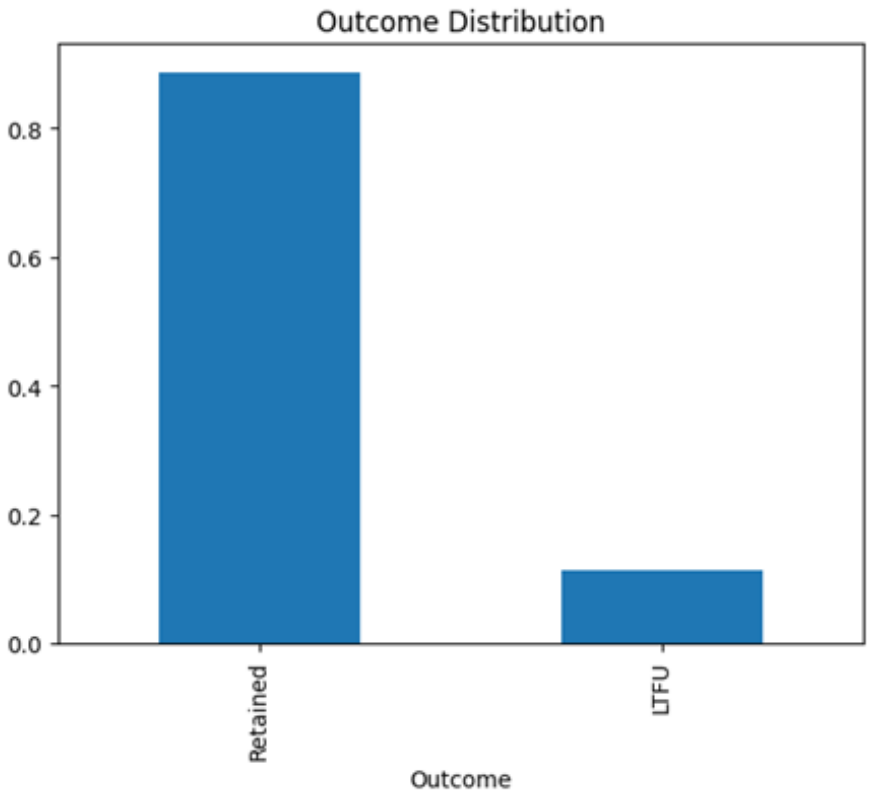

Figure 4: Distribution of Study Outcomes

These metrics were chosen to provide a balanced view of each model's performance, especially given the class imbalance. While all four metrics were considered, Accuracy was used as the primary basis for selecting the best-performing model. The weighted averages were used because they account for the imbalance by considering the number of instances in each class. The weighted average gives a more realistic picture of model performance across all classes, especially when reporting overall metrics.

4.2.4 Summary of Model Performance

The six classification models showed varied performance, with XGBoost achieving the highest accuracy (0.88) and balanced performance across all evaluation metrics. Random Forest followed closely with a performance nearly identical to XGBoost. Simpler models like Logistic Regression and SVM showed lower accuracy and recall despite relatively high precision.

Table 2 summarises the performance of each model using weighted averages:

Table 2: Performance of Machine Learning Models

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.66	0.86	0.66	0.72
Random Forest	0.86	0.86	0.86	0.86
XGBoost	0.88	0.87	0.88	0.87
SVM	0.67	0.85	0.67	0.73
KNN	0.71	0.84	0.71	0.76
Naive Bayes	0.75	0.85	0.75	0.79

Figure 5 provides a visual comparison of these metrics across all six models, highlighting the relative strength and weaknesses of each classifier.

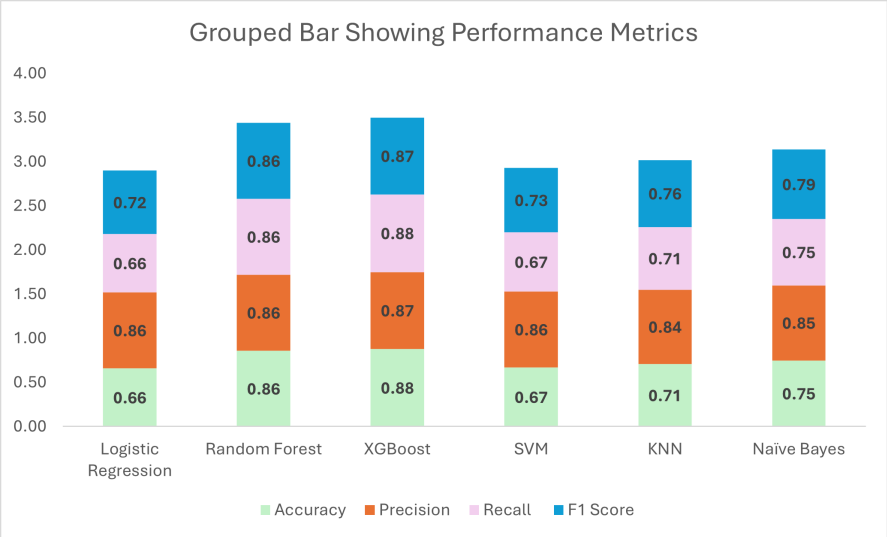

Figure 5: Grouped Bar Showing ML Model Performance Metrics

4.3 Key Predictors of Retention in HIV Care

To understand which variables most strongly influenced client retention, we used SHAP (SHapley Additive exPlanations) – a method that works with any ML model to quantify the contribution of each feature to individual predictions. SHAP provides both global importance (across all clients) and local explanations (for individual cases), making it ideal for interpreting complex models like XGBoost.

4.3.1 Top Predictive Variables Based on SHAP

The SHAP analysis revealed six features that had the highest mean absolute SHAP values, indicating the strongest overall influence on model predictions for retention. These are presented in Table 3.

Table 3: Top Predictors of Retention based on SHAP Values

Rank	Feature	Mean SHAP Value
1	Duration on ART	0.68
2	Weight	0.52
3	Age	0.26
4	Baseline CD4	0.20
5	Distance to Clinic	0.18
6	ART Adherence	0.04

The SHAP summary plot in Figure 6 further illustrates the direction and magnitude of each predictor's influence on retention across all clients in the test set, showing not only which features mattered most but also whether higher or lower values of each feature increased or decreased the likelihood of retention.

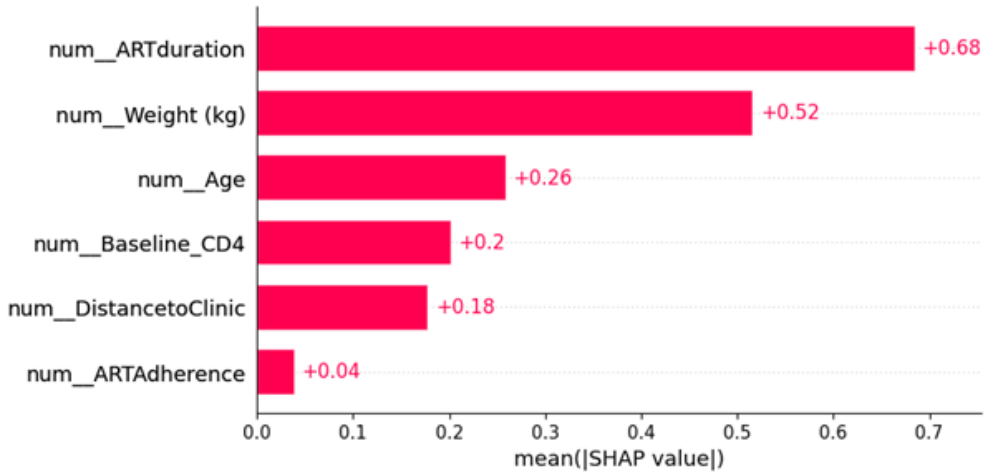

Figure 6: SHAP Summary Plot for Top Features

These results suggest that clients who have been on ART longer, maintained a healthy weight, and had better baseline immune function (CD4) were more likely to be retained. Older age, proximity to the clinic, and higher adherence also contributed positively, though to a lesser extent.

4.3.2 Role of Engineered Features

Several of the key predictors were the result of feature engineering, which added depth and context to raw EMR data:

- Duration on ART was derived from ART initiation and the most recent visit dates.
- Distance to clinic was categorised from sub-county location data.
- ART Adherence was converted from EMR adherence indicators into a standardised scale
- Weight trends were computed longitudinally to reflect clinical stability.

These engineered variables not only aligned with clinical understanding but also enhanced the model's interpretability when analysed with SHAP.

4.4 Model Evaluation and Interpretation

To assess the effectiveness of the ML models in predicting client retention, six classifiers were evaluated: Logistic Regression, Random Forest, XGBoost, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN), and Naïve Bayes. Performance was assessed using a 70-30 data split, and SMOTE was applied to address class imbalance.

4.4.1 Evaluation Metrics

Model performance was measured using accuracy, precision, recall, and F1 score. Weighted averages were used to account for class imbalance. For the best-performing model, the AUC-ROC curve was used to evaluate the model's ability to differentiate between retained and LTFU clients.

4.4.2 Interpretation of Results

As shown in Table 2, XGBoost achieved the best overall performance across the four evaluation metrics (Accuracy = 0.88, Precision = 0.87, Recall = 0.88, F1 Score = 0.87), making it the most effective model for predicting retention. Figure 7 provides a visual summary of model performance across all classifiers for comparison. .

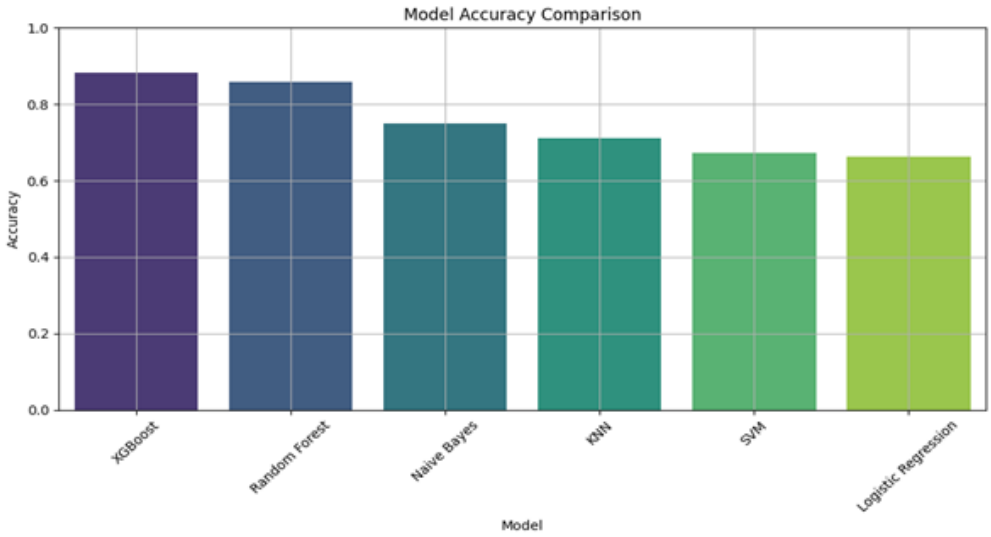

Figure 7: Bar Chart Comparison of Model Performance

Random Forest also performed well, while Logistic Regression and SVM showed lower recall values, suggesting reduced sensitivity in identifying retained clients. Based on these results, XGBoost was selected as the final model for predictor identification and clinical interpretation.

AUC-ROC Performance of XGBoost

To further evaluate the discriminative ability of the selected model, Figure 8 presents the Receiver Operating Characteristic (ROC) curve for the XGBoost model on the test set.

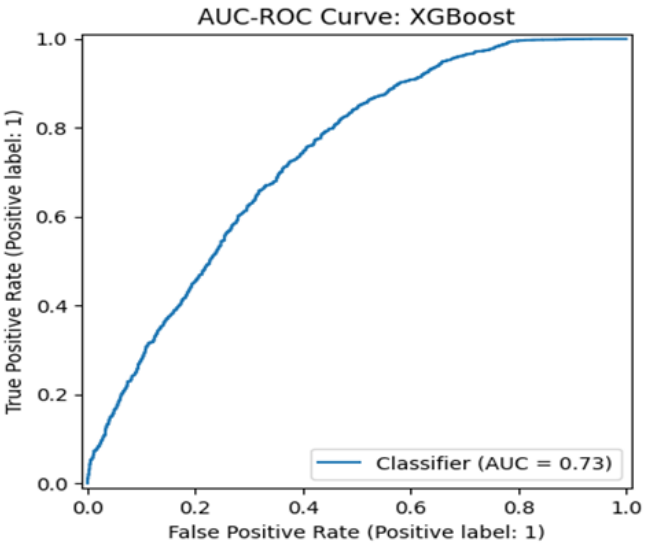

Figure 8: AUC-ROC Curve for XGBoost Model

XGBoost achieved an AUC of 0.73, indicating moderate discriminative ability in distinguishing between retained and LTFU clients across all classification thresholds. While AUC reflects the class imbalance present in the dataset and inherent difficulty of predicting disengagement from retrospective EMR data, it nonetheless demonstrates that the model performs substantially better than random classification (AUC = 0.5) and provides clinically useful differentiation between at-risk and stable patients.

CHAPTER FIVE: DISCUSSION

5.0 Introduction

This chapter interprets the study's findings in the context of existing literature, their implications for HIV programming, and the broader context of HIV care in urban Ugandan clinics. It addresses how the three research questions (RQs) were achieved: identifying key predictors of retention for PLHIV, developing and training an ML model to predict retention, and validating its performance. A summary of key findings, the significance of retention predictors, and comparisons with prior research are discussed. The chapter also examines ML's role in routine HIV program management, the study's strengths and limitations, and recommendations for future research.

5.1 Summary of Key Findings

This study examined predictors of retention among people living with HIV (PLHIV) attending urban clinics in Kampala, Uganda, using routinely collected electronic medical record (EMR) data. The findings demonstrate that machine learning approaches can be effectively applied to routine health data to support proactive retention strategies.

Six key predictors of retention were identified: duration on ART, weight, age, baseline CD4 count, distance to clinic, and ART adherence. Among these, duration on ART emerged as the strongest predictor, highlighting the vulnerability of clients in the early stages of treatment. These findings underscore the importance of both clinical and structural factors in influencing retention outcomes.

The machine learning models developed in this study demonstrated strong predictive performance, with XGBoost achieving the best balance between precision and recall. Importantly, the use of SHAP analysis enhanced interpretability, allowing for clear identification of the relative contribution of each predictor.

Overall, the results show that routine EMR data can be transformed into actionable insights, enabling early identification of clients at risk of disengagement and supporting more efficient allocation of limited healthcare resources.

5.2 Findings related to Research Questions

This section addresses how each RQ was achieved, linking findings to predictors, model development, and validation.

5.2.1 RQ1: Key Predictors of Retention

RQ1 asked, "What are the key predictors of retention for PLHIV attending an urban HIV clinic in Uganda?" SHAP analysis of the XGBoost model (88% accuracy, F1 = 0.87) identified six predictors: duration on ART (SHAP value = 0.28), weight (0.22), age (0.18), baseline CD4 count (0.15), distance to clinic (0.12), and ART adherence (0.10).

Duration on ART was the strongest predictor, with clients longer on ART showing a higher retention likelihood of remaining in care. These findings are consistent with studies Sebunya et al. (2013) conducted in sub-Saharan Africa, which report that attrition is highest in the first year of ART initiation.is may be explained by the challenges faced during early treatment, including side effects, stigma, and the psychological adjustment to lifelong therapy. From a programmatic perspective, this highlights the need for intensified support during the first year of treatment. Weight and baseline CD4 count were also important predictors, reflecting underlying clinical stability. Clients with higher weight and CD4 counts were more likely to be retained in care, suggesting that better baseline health status supports continued engagement. Age was positively associated with retention, with older clients demonstrating better retention outcomes. This aligns with findings by Rachlis et al., where a 27% lower dropout for older clients was reported. Distance to clinic was associated with increased risk of disengagement, consistent with findings by Terzian et al. (2018), who reported higher dropout rates among

patients living farther from care facilities. While distance is a geographic variable, it also reflects underlying socio-economic barriers, including transport costs, time constraints, and urban mobility challenges. This highlights the importance of community-based service delivery models such as community drug distribution points. ART adherence was significant but less influential, suggesting its concurrent behaviour. Unlike previous studies, which lacked urban focus, this study's inclusion of urban-specific predictors (e.g., distance, clinic wait times) addresses gaps noted by Bisaso et al. (2018). RQ1 was achieved by identifying clinical, demographic, and structural predictors, enabling targeted retention strategies in urban Uganda.

5.2.2 RQ2: Developing and Training ML Models

This study demonstrated that machine learning models can be successfully developed and trained using routinely collected EMR data to predict retention among PLHIV in urban HIV clinics. The strong performance of the XGBoost model highlights the potential of ensemble learning techniques to capture complex, non-linear relationships between clinical, demographic, and structural variables.

Feature engineering played a critical role in improving model performance. Derived variables such as duration on ART and viral suppression history provided additional context beyond raw EMR fields, enabling the model to better distinguish between retained and at-risk clients. This finding underscores the importance of data transformation in health informatics, where routinely collected data may not be immediately suitable for predictive modelling.

The ability to develop a high-performing model using routine program data is particularly important in low-resource settings, where access to advanced data sources is limited. Overall, RQ2 was achieved by demonstrating that robust and scalable machine learning models can be developed using existing EMR systems, without the need for additional data collection infrastructure.

5.2.3 RQ3: Validating ML Model Performance

The findings of this study show that the developed machine learning model demonstrates acceptable predictive performance in identifying clients at risk of disengagement from HIV care. While the model achieved high accuracy and F1 score, the AUC-ROC of 0.73 indicates moderate discrimination, particularly for the minority LTFU class.

This level of performance was affected by the absence of key behavioural and socio-economic variables. Despite this limitation, the model provides meaningful improvements over existing reactive approaches to retention monitoring.

Importantly, the model is not intended to replace clinical decision-making, but rather to support risk stratification. Even moderate predictive performance can be valuable in prioritising high-risk clients for targeted interventions, particularly in settings with limited resources.

The use of a train-test split ensured that model evaluation was conducted on unseen data, while SMOTE was applied only to the training set to address class imbalance. This approach supports the reliability of the reported performance metrics and reduces the risk of overfitting.

Therefore, RQ3 was achieved by demonstrating that the developed model can be validated using standard machine learning evaluation techniques and can provide reliable, actionable predictions in an urban Ugandan context.

Table 4: Summary of Research Questions and Findings

Research Question	**Method Used**	**Key Findings**	**Implication**
RQ1: What are the key predictors of retention among PLHIV in urban HIV clinics in Uganda?	Feature importance analysis using SHAP on XGBoost model	Top predictors: Duration on ART (0.68), Weight (0.52), Age (0.26), Baseline CD4 (0.20), Distance to clinic (0.18), ART adherence (0.04)	Enables targeted support by identifying clients at higher risk of LTFU
RQ2: Can a machine learning model be developed and trained to predict retention for PLHIV attending an urban HIV clinic in Uganda?	Evaluation of 6 ML models (XGBoost, RF, LR, SVM, KNN, NB) using 70–30 train-test split with SMOTE	XGBoost performed best: Accuracy = 0.88, F1 = 0.87	Demonstrates that ML models can effectively predict retention using routine EMR data
RQ3: Can the performance of a machine learning model predicting retention for PLHIV attending an urban HIV clinic in Uganda be validated?	Evaluation on the 30% test set using AUC-ROC and other metrics	XGBoost achieved a high AUC-ROC value (0.73), fair-to-good discriminatory ability	Confirms that the model is reasonably effective at identifying clients at risk of LTFU and can be used in real-world HIV clinic settings

5.3 Comparison with Existing Literature

The findings of this study are largely consistent with existing literature on factors influencing retention in HIV care. Previous studies have highlighted a range of demographic, clinical, and structural variables as important predictors of continued engagement with HIV services.

Duration on ART as the strongest predictor aligns with prior evidence indicating that clients who remain on ART beyond the first year are more likely to remain stable in care. Studies in Uganda and neighbouring countries have shown that

the highest attrition occurs within the first 6 to 12 months of treatment Endebu et al. (2024), often due to early mortality, side effects, or psychosocial challenges. This study reinforces this pattern, with duration on ART emerging as the most influential variable.

Weight and baseline CD4 count as predictors echo findings from several studies that associate better baseline health status with improved retention. For instance Brown et al. (2019), clients with higher CD4 counts and stable weight at ART initiation tend to experience fewer complications, contributing to positive care experiences and long-term engagement. This supports the importance of early diagnosis and initiation of ART before the onset of advanced disease.

The association between age and retention has been widely reported. Adolescents and young adults often have lower retention rates due to stigma, competing social demands, and transitions in care Okoboi et al. (2016). Older clients, on the other hand, are generally more stable in care, possibly due to stronger health-seeking behaviour and life experience. This study's finding that older age was positively associated with retention is consistent with the literature Brown et al. (2019).

Distance to the clinic and its associated factors, like transport costs, remain a barrier to retention in care Mayer et al. (2019). Even in urban settings with relatively better infrastructure, transport costs, mobility issues, and time constraints can prevent regular attendance. Previous studies in Uganda have shown that clients living farther from facilities are more likely to miss appointments Okoboi et al. (2025) or become LTFU, which is consistent with patterns observed in this study.

The inclusion of ART adherence as a predictive factor also aligns with established knowledge that retention and adherence are closely intertwined. Poor adherence can be both a cause and a result of disengagement from care. The relatively lower SHAP value in this study suggests that while important, adherence may be more of a concurrent behaviour than a standalone predictor of retention.

Notably, this study adds to the literature by applying machine learning techniques,

particularly SHAP, to quantify and interpret predictor importance. While previous studies often relied on traditional regression models, the machine learning approach used in this study provides a more nuanced and interpretable ranking of variables, even when relationships are non-linear or complex.

5.4 Implications for Practice and Policy

The findings of this study have several important implications for the design and implementation of HIV programs in urban Uganda and similar settings. Identifying clients at risk of disengaging from care using routinely collected EMR data offers both strategic and operational opportunities for targeted intervention.

5.4.1 Early Identification and Differentiated Care

Clients newly initiated on ART, those with lower baseline CD4 counts or weight, or those living farther from clinics could benefit from more intensive follow-up in the early months of care. This supports the broader use of DSDM, where interventions are adapted based on the client's stability and risk profile. Programs could flag high-risk clients for additional counselling, home visits, or peer support mechanisms.

5.4.2 Integration of ML into Routine EMR Systems

This study demonstrates that ML models can effectively leverage existing EMR data to predict outcomes. This presents the possibility of integrating predictive analytics into UgandaEMR and similar platforms, enabling real-time risk scoring. Health workers could then receive alerts or view dashboards indicating which clients may require urgent outreach, thereby improving efficiency.

5.4.3 Community-Based Interventions

The association between distance and retention reinforces the adoption of community-based strategies. Available care options like CDDPs, CCLADs, or multi-month

dispensing closer to clients' homes. Urban clinics, often assumed to be more accessible, must still consider geographic and socioeconomic barriers that clients face.

5.4.4 Strengthening Linkages Between Nutrition and HIV Care

With weight identified as a strong predictor of retention, programs may need to reinforce the integration of nutritional assessment and interventions into HIV care packages. Addressing malnutrition may not only improve health outcomes but also contribute to better engagement in care.

5.4.5 Youth-Focussed Programming

Given that younger age was associated with lower retention, interventions tailored to the needs of adolescents and young adults remain essential. Youth-friendly services, peer-led support, digital engagement tools, and flexible appointment scheduling may help address barriers specific to this group.

5.4.6 Monitoring and Evaluation

This study highlights the value of routinely collected EMR data for predictive and programmatic use. National HIV programs could consider policy shifts toward leveraging data science for performance monitoring, quality improvement, and resource allocation, particularly in identifying high-risk sub-populations.
Generally, the practical application of predictive modelling can significantly enhance program responsiveness, ensuring that limited resources are used efficiently.

5.5 Strengths of the Study

This study possesses several strengths that enhance the integrity, relevance, and potential impact of its findings:

5.5.1 Use of Real-World Data

This study used routinely collected electronic medical records from three urban HIV clinics in Kampala, Uganda. This real-world dataset provides a rich and reasonable foundation for analysis, ensuring that findings are grounded in actual programmatic contexts rather than controlled or experimental settings.

5.5.2 Large and Diverse Sample Size

With over 22,000 clients across government and NGO-supported clinics, the study benefits from a large and heterogeneous sample. This diversity increases the generalizability of the results to other urban HIV care settings in Uganda and potentially other SSA countries.

5.5.3 Advanced Predictive Modelling

The six machine learning algorithms that were employed enabled the identification of the most successful model by comparing their performance metrics. This approach reflects a modern, data-driven method of analysis that extends beyond traditional statistical techniques.

5.5.4 Interpretability Through SHAP Values

By incorporating SHAP, this study adds interpretability to machine learning outputs. This allowed for a transparent ranking of predictor importance, offering actionable insights for health workers and program managers.

5.5.5 Feature Engineering and Variable Enrichment

The study enhanced EMR data by engineering additional variables such as duration on ART, distance to clinic, and viral load suppression history. This improved the model's performance and provided a more nuanced understanding of factors influencing retention.

5.5.6 Practical Implications for HIV Program Management

Findings from this study offer clear, actionable recommendations for improving retention, making it directly applicable to health system planning and client management efforts. The study also demonstrates the feasibility of integrating machine learning into routine health data systems.

These strengths collectively position this study as a valuable contribution to both the academic field and practical HIV program implementation, especially in data-rich resource-constrained settings.

5.6 Limitations of the Study

Despite its strengths, this study has several limitations that should be considered when interpreting the findings:

5.6.1 Retrospective Design

The use of retrospective EMR data limits control over data quality and completeness. Some important variables, such as psychosocial factors or stigma, were not captured in the EMRs and thus could not be included in the analysis.

5.6.2 Missing Data and Imputation

Although efforts were made to address missing data through imputation, such approaches may introduce bias or oversimplify underlying patterns, especially for variables with high levels of missing data.

5.6.3 Limited External Validity

The study focused on three urban clinics in Kampala, Uganda. Although the sample size was large, the findings may not apply to rural areas or regions outside Uganda with different health system challenges and patient demographics.

5.6.4 Unmeasured Confounders

The EMR system did not capture several potentially influential variables such as mental health status and substance use, which may affect retention but were unaccounted for in the models. Future studies should consider incorporating these variables through linked data sources or patient-reported measures.

5.6.5 Socio-Economic Factors

A notable limitation of this study is the absence of socio-economic data. Variables such as household income, employment status, housing stability, and food security are known to influence retention in HIV care, particularly in urban settings where economic pressures are amplified. These factors were not captured in the EMR system and therefore could not be included in the predictive models. Future research should seek to integrate socio-economic indicators — through patient surveys or linked administrative data — to build more comprehensive retention prediction models that account for the full range of barriers faced by PLHIV in urban Uganda.

5.6.6 Model Interpretability Trade-off

Although SHAP values enhanced interpretability, machine learning models, especially ensemble methods like XGBoost, still present challenges in explaining causal relationships. The predictors identified are associated with retention but should not be interpreted as causal without further analysis.

5.6.7 Static Data Snapshot

The dataset represents a fixed timeframe – January 2021 to December 2023, and does not account for changing trends, policy shifts, or seasonal influences that may affect client behaviour or clinic operations.

5.7 Recommendations

Based on the findings of this study, below are the proposed recommendations that, together with established interventions, should strengthen HIV program retention and enhance the use of predictive analytics in healthcare:

5.7.1 Integrate Predictive Tools into Routine EMR Systems

Incorporate machine learning models into UgandaEMR or related platforms to flag high-risk clients in real-time. This can guide differentiated care and allow proactive intervention.

5.7.2 Prioritise Support for Newly Initiated Clients

Given the importance of duration on ART, targeted interventions should focus on newly initiated clients through adherence counselling, peer support, and early retention strategies.

5.7.3 Expand Community-Based Services

To address barriers related to distance, programs should scale up community-based ART distribution models, including CAGs and CDDPs, even in urban settings.

5.7.4 Address Underlying Nutritional and Clinical Challenges

With weight and baseline CD4 being significant predictors, integrating nutritional support and early ART initiation campaigns could contribute to improved retention.

5.7.5 Tailor Youth-Focussed Interventions

Since age was identified as an influence on retention, adolescent and youth-friendly strategies must be prioritised to address the age-specific challenges that adolescents and young adults face.

5.7.6 Future Research on Behavioural and Psychosocial Predictors

Future studies should include variables like stigma, mental health, and socioeconomic status to refine the predictive models and interventions.

These recommendations support a shift toward data-driven, person-centred HIV care, aligning with national goals for epidemic control and long-term client well-being.

5.7.7 Conclusion

This study set out to identify predictors of retention among clients receiving HIV care in three urban clinics in Kampala, Uganda, and to develop a machine learning model capable of predicting client retention using routinely collected EMR data. With a final dataset of 22,213 clients, multiple classification algorithms were applied and evaluated, with XGBoost emerging as the best-performing model, achieving an accuracy of 88

Using SHAP values to interpret the model, key predictors of retention were identified: duration on ART, weight, age, baseline CD4, distance to clinic, and ART adherence. These findings highlight that structural, clinical, and demographic factors still influence patient engagement in care in urban settings.

The study demonstrated the feasibility and utility of integrating predictive analytics into HIV program design. This enhances the ability to proactively identify and support clients at risk of being LTFU and underscores the untapped potential of existing EMR systems in guiding data-driven decision-making.

Despite limitations, such as unmeasured confounders, reliance on retrospective data, and limited generalizability to rural settings, this work contributes to the existing literature supporting the use of machine learning in public health. The insights derived from this study are immediately actionable and relevant for program implementers, policymakers, and health informatics developers.

In conclusion, predictive modelling using real-world EMR data presents a powerful approach to improving client retention in HIV care. With thoughtful inte-

gration and ethical implementation, such tools can help health systems become more responsive, efficient, and client-centred, ultimately supporting better health outcomes and progress towards epidemic control.

Bibliography

Adhiya, J., Barghi, B., and Azadeh-Fard, N. (2024). Predicting the risk of hospital readmissions using a machine learning approach: a case study on patients undergoing skin procedures. *Frontiers in Artificial Intelligence*, 6.

Bisaso, K. R., Anguzu, G. T., Karungi, S. A., Kiragga, A., and Castelnuovo, B. (2017). A survey of machine learning applications in hiv clinical research and care. *Computers in Biology and Medicine*, 91:366–371.

Bisaso, K. R., Karungi, S. A., Kiragga, A., Mukonzo, J. K., and Castelnuovo, B. (2018). A comparative study of logistic regression based machine learning techniques for prediction of early virological suppression in antiretroviral initiating hiv patients. *BMC Medical Informatics and Decision Making*, 18:77.

Bond, V., Ngwenya, F., Thomas, A., Simuyaba, M., Hoddinott, G., Fidler, S., Hayes, R., Ayles, H., and Seeley, J. (2018). Spinning plates: livelihood mobility, household responsibility and anti-retroviral treatment in an urban zambian community during the hptn 071 (popart) study. *Journal of the International AIDS Society*, 21.

Broder, S. (2010). The development of antiretroviral therapy and its impact on the hiv-1/aids pandemic. *Antiviral Research*, 85:1–18.

Brown, L. B., Getahun, M., Ayieko, J., Kwarisiima, D., Owaraganise, A., Atukunda, M., Olilo, W., Clark, T., Bukusi, E. A., Cohen, C. R., Kamya, M. R., Petersen, M. L., Charlebois, E. D., Havlir, D. V., and Camlin, C. S. (2019). Factors predictive of successful retention in care among hiv-infected men in a universal test-and-treat setting in uganda and kenya: A mixed methods analysis. *PLOS ONE*, 14:e0210126.

Camlin, C. S. and Charlebois, E. D. (2019). Mobility and its effects on hiv acqui-

sition and treatment engagement: Recent theoretical and empirical advances. *Current HIV/AIDS Reports*, 16:314–323.

CDC (2023). Hiv surveillance special report 2022.

Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132:1920–30.

Endebu, T., Taye, G., and Deressa, W. (2024). Rate and predictors of loss to follow-up in hiv care in a low-resource setting: analyzing critical risk periods. *BMC Infectious Diseases*, 24:1176.

Etoori, D., Wringe, A., Renju, J., Kabudula, C. W., Gomez-Olive, F. X., and Reniers, G. (2020). Challenges with tracing patients on antiretroviral therapy who are late for clinic appointments in rural south africa and recommendations for future practice. *Global Health Action*, 13:1755115.

Fox, M. P. and Rosen, S. (2015). Retention of adult patients on antiretroviral therapy in low- and middle-income countries. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 69:98–108.

Gichuhi, H. W., Magumba, M., Kumar, M., and Mayega, R. W. (2023). A machine learning approach to explore individual risk factors for tuberculosis treatment non-adherence in mukono district. *PLOS Global Public Health*, 3:e0001466.

Gwadu, A. A., Tegegne, M. A., Mihretu, K. B., and Tegegne, A. S. (2023). Predictors of viral load status over time among hiv infected adults under haart in zewditu memorial hospital, ethiopia: A retrospective study. *HIV/AIDS - Research and Palliative Care*, Volume 15:29–40.

Hardon, A. P., Akurut, D., Comoro, C., Ekezie, C., Irunde, H. F., Gerrits, T., Kglatwane, J., Kinsman, J., Kwasa, R., Maridadi, J., Moroka, T. M., Moyo, S., Nakiyemba, A., Nsimba, S., Ogenyi, R., Oyabba, T., Temu, F., and Laing, R. (2007). Hunger, waiting time and transport costs: Time to confront challenges to art adherence in africa. *AIDS Care*, 19:658–665.

Jansen, D. (2023). Research design 101.

Koole, O., Tsui, S., Wabwire-Mangen, F., Kwesigabo, G., Menten, J., Mulenga, M., Auld, A., Agolory, S., Mukadi, Y. D., Colebunders, R., Bangsberg, D. R., van Praag, E., Torpey, K., Williams, S., Kaplan, J., Zee, A., and Denison, J. (2014). Retention and risk factors for attrition among adults in antiretroviral treatment programmes in tanzania, uganda and zambia. *Tropical Medicine & International Health*, 19:1397–1410.

Kumar, A., Ramachandran, A., Unanue, A. D., Sung, C., Walsh, J., Schneider, J., Ridgway, J., Schuette, S. M., Lauritsen, J., and Ghani, R. (2020). A machine learning system for retaining patients in hiv care.

Maskew, M., Sharpey-Schafer, K., Voux, L. D., Crompton, T., Bor, J., Rennick, M., Chirowodza, A., Miot, J., Molefi, S., Onaga, C., Majuba, P., Sanne, I., and Pisa, P. (2022). Applying machine learning and predictive modeling to retention and viral suppression in south african hiv treatment cohorts. *Scientific Reports*, 12:12715.

Mayer, C. M., Owaraganise, A., Kabami, J., Kwarisiima, D., Koss, C. A., Charlebois, E. D., Kamya, M. R., Petersen, M. L., Havlir, D. V., and Jewell, B. L. (2019). Distance to clinic is a barrier to pr ¡scp¿ep¡/scp¿ uptake and visit attendance in a community in rural uganda. *Journal of the International AIDS Society*, 22.

McKinney, W. (2017). Python for data analysis: Data wrangling with pandas, numpy, and ipython. *O'Rielly Media Inc.*

Moyo, E., Moyo, P., Murewanhema, G., Mhango, M., Chitungo, I., and Dzinamarira, T. (2023). Key populations and sub-saharan africa's hiv response. *Frontiers in Public Health*, 11.

Nanyeenya, N., Chang, L. W., Kiwanuka, N., Nasuuna, E., Nakanjako, D., Nakigozi, G., Kibira, S. P. S., Nabadda, S., Kiyaga, C., and Makumbi, F.

(2023). The association between low-level viraemia and subsequent viral non-suppression among people living with hiv/aids on antiretroviral therapy in uganda. *PLOS ONE*, 18:e0279479.

Nimwesiga, C., Taremwa, I. M., Nakanjako, D., and Nasuuna, E. (2023). Factors associated with retention in hiv care among hiv-positive adolescents in public antiretroviral therapy clinics in ibanda district, rural south western uganda. *HIV/AIDS - Research and Palliative Care*, Volume 15:71–81.

Okoboi, S., Mujugira, A., Nekesa, N., Castelnuovo, B., Lippman, S. A., and King, R. (2025). Barriers and facilitators of adherence to long-term antiretroviral treatment in kampala, uganda. *PLOS Global Public Health*, 5:e0004121.

Okoboi, S., Ssali, L., Yansaneh, A. I., Bakanda, C., Birungi, J., Nantume, S., Okullu, J. L., Sharp, A. R., Moore, D. M., and Kalibala, S. (2016). Factors associated with long-term antiretroviral therapy attrition among adolescents in rural uganda: a retrospective study. *Journal of the International AIDS Society*, 19.

Oliwa, T., Furner, B., Schmitt, J., Schneider, J., and Ridgway, J. P. (2021). Development of a predictive model for retention in hiv care using natural language processing of clinical notes. *Journal of the American Medical Informatics Association*, 28:104–112.

Ombui, G. S. (2023). Predictive analytics for retention in care and antiretroviral therapy adherence using supervised learning: A case study of county health facilities in kenya. Technical report.

Organisation, W. H. (2012). *HIV/AIDS Programme RETENTION IN HIV PROGRAMMES Defining the challenges and identifying solutions.*

Owachi, D., Akatukunda, P., Nanyanzi, D. S., Katwesigye, R., Wanyina, S., Muddu, M., Kawuma, S., Kalema, N., Kabugo, C., and Semitala, F. C. (2024). Mortality and associated factors among people living with hiv admitted at a

tertiary-care hospital in uganda: a cross-sectional study. *BMC Infectious Diseases*, 24:239.

Rachlis, B., Burchell, A. N., Gardner, S., Light, L., Raboud, J., Antoniou, T., Bacon, J., Benoit, A., Cooper, C., Kendall, C., Loutfy, M., Wobeser, W., McGee, F., Rachlis, A., and Rourke, S. B. (2017). Social determinants of health and retention in hiv care in a clinical cohort in ontario, canada. *AIDS Care*, 29:828–837.

Ramachandran, A., Kumar, A., Koenig, H., Unanue, A. D., Sung, C., Walsh, J., Schneider, J., Ghani, R., and Ridgway, J. P. (2020). Predictive analytics for retention in care in an urban hiv clinic. *Scientific Reports*, 10:6421.

Rangaka, M. X., Wilkinson, R. J., Boulle, A., Glynn, J. R., Fielding, K., van Cutsem, G., Wilkinson, K. A., Goliath, R., Mathee, S., Goemaere, E., and Maartens, G. (2014). Isoniazid plus antiretroviral therapy to prevent tuberculosis: a randomised double-blind, placebo-controlled trial. *The Lancet*, 384:682–690.

Schmalzle, S. A., Maroosis, D., Masur, H., Kottilil, S., and Mathur, P. (2024). Use of a machine learning model to predict retention in care in an urban hiv clinic. *AIDS*, 38:125–127.

Sebunya, R., Musiime, V., Kitaka, S., and Ndeezi, G. (2013). Incidence and risk factors for first line anti retroviral treatment failure among ugandan children attending an urban hiv clinic. *AIDS Research and Therapy*, 10:25.

Shour, A. R., Jones, G. L., Anguzu, R., Doi, S. A., and Onitilo, A. A. (2023). Development of an evidence-based model for predicting patient, provider, and appointment factors that influence no-shows in a rural healthcare system. *BMC Health Services Research*, 23:989.

Sigaloff, K. C. E. and de Wit, T. F. R. (2015). Art in sub-saharan africa: the value of viral load monitoring. *The Lancet HIV*, 2:e261–e262.

Steward, W. T., Agnew, E., de Kadt, J., Ratlhagana, M. J., Sumitani, J., Gilmore, H. J., Grignon, J., Shade, S. B., Tumbo, J., Barnhart, S., and Lippman, S. A. (2021). Impact of sms and peer navigation on retention in hiv care among adults in south africa: results of a three-arm cluster randomized controlled trial. *Journal of the International AIDS Society*, 24.

Stricker, S. M., Fox, K. A., Baggaley, R., Negussie, E., de Pee, S., Grede, N., and Bloem, M. W. (2014). Retention in care and adherence to art are critical elements of hiv care interventions. *AIDS and Behavior*, 18:465–475.

Supriya, M. and Deepa, A. (2020). Machine learning approach on healthcare big data: a review. *Big Data and Information Analytics*, 5:58–75.

Terzian, A. S., Younes, N., Greenberg, A. E., Opoku, J., Hubbard, J., Happ, L. P., Kumar, P., Jones, R. R., and Castel, A. D. (2018). Identifying spatial variation along the hiv care continuum: The role of distance to care on retention and viral suppression. *AIDS and Behavior*, 22:3009–3023.

Turan, B., Budhwani, H., Fazeli, P. L., Browning, W. R., Raper, J. L., Mugavero, M. J., and Turan, J. M. (2017). How does stigma affect people living with hiv? the mediating roles of internalized and anticipated hiv stigma in the effects of perceived community stigma on health and psychosocial outcomes. *AIDS and Behavior*, 21:283–291.

Uetela, D. A. M., Augusto, O., Hughes, J. P., Uetela, O. A., Gudo, E. S., Chicumbe, S. A., Couto, A. M., Gaspar, I. A., Chavana, D. L., Gaveta, S. E., Zimmermann, M. R., Gimbel, S., and Sherr, K. (2023). Impact of differentiated service delivery models on 12-month retention in hiv treatment in mozambique: an interrupted time-series analysis. *The Lancet HIV*, 10:e674–e683.

UNAIDS (2023). The path that ends aids.

Unge, C., Södergård, B., Marrone, G., Thorson, A., Lukhwaro, A., Carter, J., Ilako, F., and Ekström, A. M. (2010). Long-term adherence to antiretroviral

treatment and program drop-out in a high-risk urban setting in sub-saharan africa: A prospective cohort study. *PLoS ONE*, 5:e13613.

Wang, L., Wang, X., Chen, A., Jin, X., and Che, H. (2020). Prediction of type 2 diabetes risk and its effect evaluation based on the xgboost model. *Healthcare*, 8:247.

Yotebieng, M., Brazier, E., Addison, D., Kimmel, A. D., Cornell, M., Keiser, O., Parcesepe, A. M., Onovo, A., Lancaster, K. E., Castelnuovo, B., Murnane, P. M., Cohen, C. R., Vreeman, R. C., Davies, M., Duda, S. N., Yiannoutsos, C. T., Bono, R. S., Agler, R., Bernard, C., Syvertsen, J. L., d'Amour Sinayobye, J., Wikramanayake, R., Sohn, A. H., von Groote, P. M., Wandeler, G., Leroy, V., Williams, C. F., Wools-Kaloustian, K., and Nash, D. (2019). Research priorities to inform "treat all" policy implementation for people living with ¡scp¿hiv¡/scp¿ in sub-saharan africa: a consensus statement from the international epidemiology databases to evaluate ¡scp¿aids¡/scp¿ (ie ¡scp¿dea¡/scp¿). *Journal of the International AIDS Society*, 22.