FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 10 No. 23, December, 2026, pp 229 – 235
DOI: https://doi.org/10.33003/fjs-2026-1023-6512

Modeling the Survival Chance of HIV Positive Patients Using Logistic Regression And Distributed Fp-Tree Algorithm

***Mohammed Umar Mohammed, Abraham Okolo and Adesupo Adeoye Akinrefon**

Department of Statistics, Modibbo Adama University, Yola, Adamawa State, Nigeria

* Corresponding authors' email: mohammedumar1022@gmail.com Phone: +2348147512798

Received: 31 Sept. 2026
Accepted: 10 Nov. 2026
Published: 16 Nov. 2026

ABSTRACT

This retrospective study examined factors associated with vital status among 200 HIV-positive patients receiving antiretroviral therapy (ART) at Specialist Hospital, Gombe, Nigeria. Patients were selected by systematic sampling from ART clinic registers. The outcome was vital status as recorded in the register, coded as alive (1) or dead (0); 162 patients (81.0%) were alive and 38 patients (19.0%) were dead. Follow-up duration and dates of records were not available in the extracted dataset, so time-to-event analysis was not possible and vital status was analysed as a binary outcome. Logistic regression was used to estimate odds ratios (OR) with 95% confidence intervals (CI), and distributed FP-Growth association rule mining (minimum support 10%, minimum confidence 60%, lift >1.0) was used to identify frequent combinations of clinical characteristics associated with death. Distributed FP-Growth was implemented using Apache Spark MLlib FPGrowth (Spark 3.5, PySpark 3.5). At n = 200, distributed runtime was 0.72 ms versus serial 0.72 ms (speedup 0.99), showing that partitioning overhead dominates at this scale. The strong association for CD4 count below 200 cells/μL reflects the near-complete separation in this dataset: 35 of 41 patients with CD4 below 200 died, while only 3 of 159 patients with CD4 of 200 or more died Ethical approval was obtained. CD4 count below 200 cells/μL was associated with lower odds of being alive (OR = 0.01, 95% CI = 0.001 to 0.08, p < 0.001) and advanced clinical stage (Stage 3-4 versus Stage 1-2) was associated with lower odds of being alive (OR = 0.13, 95% CI = 0.02 to 0.82, p = 0.029), tuberculosis comorbidity (OR = 8.55, 95% CI = 1.39 to 52.51, p = 0.021). Age, viral load, adherence, and were not robustly associated. The model had apparent AUC 0.971, sensitivity 0.975, specificity 0.921, Brier score 0.036, and 10-fold cross-validated AUC 0.908 and CV Brier score 0.046. Distributed FP-Growth identified frequent co-occurrence of advanced stage, CD4 <200 cells/μL, and poor adherence among patients who died. Findings are associational and require external validation.

Keywords: HIV, Antiretroviral Therapy, Logistic Regression, Distributed FP-Growth, Apache Spark, Treatment Outcomes

INTRODUCTION

HIV continues to impose a substantial burden on health systems in sub-Saharan Africa, and Nigeria accounts for a large share of people living with HIV in the region. Although antiretroviral therapy has substantially improved survival, a proportion of patients still experience adverse outcomes, including death and treatment failure. HIV continues to be a significant global health challenge, particularly in low and middle income countries. Waziri et al. (2015), UNAIDS (2019) and WHO (2022) reports revealed that sub-Saharan Africa remains the most affected region, accounting for over two-thirds of global HIV cases. Survival among HIV-positive patients is influenced by multiple factors, including immune system status (CD4 count), viral load, treatment adherence, co-infections, age, and socio-economic conditions (Bisson et al., 2017; Pantaleo & Fauci, 2019; Emmanuel et al., 2023). Several studies have emphasized the importance of early initiation of antiretroviral therapy (ART) in improving survival outcomes. For example, the START trial demonstrated that early ART significantly reduces the risk of AIDS-related events and mortality (INSIGHT START Study Group, 2015). Mills et al. (2019) highlighted that treatment adherence is a critical determinant of survival, with poor adherence leading to higher viral load, treatment failure, and increased mortality. Sterne et al. (2016) reported that comorbidities such as tuberculosis and hepatitis C virus (HCV) co-infections further complicate survival outcomes, emphasizing the need for comprehensive risk assessment. Bansi et al. (2018) documented that demographic factors such as age, gender, and socio-economic status also influence survival probability, with older patients and those with limited access to healthcare resources often exhibiting poorer outcomes.

Survival prediction is critical for clinical management because it allows healthcare providers to identify high-risk patients and implement timely interventions (Karcher et al., 2007; Martha, 2012). Traditional methods of assessing prognosis, such as clinician judgment and single-variable analysis, may not fully capture complex interactions among risk factors (Kilbourne et al., 2002; Nachega, 2005; Bisson et al., 2017). Logistic regression is one of the most commonly used statistical methods for modeling binary outcomes, such as survival versus mortality (Hosmer et al., 2013). It estimates the probability of an event occurring as a function of one or more predictor variables. In HIV research, logistic regression has been widely applied to identify risk factors influencing survival and to quantify the effect of clinical and demographic variables. However, logistic regression has limitations in capturing complex interactions between multiple variables and non-linear relationships, which may reduce predictive accuracy (Steyerberg, 2019).

Recent research in healthcare have explored hybrid approach that combines data mining and statistical techniques to improve survival prediction accuracy. This approach leverages the interpretability of regression models while incorporating insights from patterns or feature interactions discovered through data mining (Zhang, 2020). This study integrates logistic regression with FP-Tree algorithm, a data mining technique, to develop a predictive model that does not only estimate survival probability but also uncovers meaningful patterns in patient characteristics. This hybrid approach offers a more comprehensive understanding of factors influencing survival among HIV-positive patients.

Logistic regression is well suited to quantifying the independent association of each factor with the odds of an outcome while adjusting for other factors. Association rule mining, in contrast, is well suited to identifying frequent combinations of characteristics that co-occur among patients with a given outcome. FP-Growth is an efficient association rule mining algorithm that identifies frequent itemsets without candidate generation. As HIV datasets grow in size and dimensionality, serial FP-Growth becomes computationally expensive. Distributed FP-Growth partitions the transaction database across workers, counts frequent itemsets in parallel, and mines local FP-Trees in parallel, which improves scalability while returning the same frequent itemsets as the serial algorithm because FP-Growth is deterministic. This study therefore combined logistic regression and distributed FP-Growth to examine both individual and combined correlates of vital status among HIV-positive patients receiving ART in Gombe, Nigeria.

Related Review

Kassaw et al. (2024) meta-analysed 15 studies (n = 5,210) of TB-HIV patients in Ethiopia and reported pooled mortality 18.73% and incidence 4.94 per 100 person-months, with CD4 <200, advanced stage, poor adherence, and lack of preventive therapy as predictors. Strength: pooled estimate across settings. Limitation: heterogeneity, no patient-level combination analysis, and no rule mining.

Wolelaw et al. (2025) followed 320 TB-HIV patients in Gambella (2017-2022) using Cox regression and reported 14.7% mortality, with female sex, poor adherence, non-use of cotrimoxazole, and viral load >1000 as predictors. Strength: time-to-event design with incidence rate. Limitation: single region, TB-only population, cannot generalise to all HIV patients.

A 2024-2025 Ethiopian study in Horro Guduru Wallaga (n = 538) used Kaplan-Meier and Cox and found 8% died, incidence 2.81 per 1,000 person-months, with WHO Stages III-IV (AHR 7.02) and failure to take TB prophylaxis (AHR 5.61) as predictors. Strength: clear time-to-event analysis. Limitation: no validation, no combination patterns.

Meloni et al. (2016) evaluated 70,002 patients and found lower CD4, higher viral load, and inadequate adherence predicted virological failure, while immunologic failure had slightly different predictors. Strength: very large sample, multi-site. Limitation: high loss to follow-up (30.1%), and immunologic criteria poorly predicted virologic failure showing that a single marker is insufficient.

Emmanuel et al. (2023) examined factors associated with HIV/AIDs in Adamawa State, Nigeria using Log-linear Model. For this study, secondary data was collected from a total population of 3,779 patients infected with HIV/AIDs who came for treatment at Specialist Hospital, Yola, Adamawa State, Nigeria. Information on gender, age, viral load, marital status and year of occurrence of the disease were extracted from their medical records. The R programming software was used to analyze the data. Fourteen (14) log linear models were fitted to the data where the model on independence of marital status, age, year, gender, viral load and interaction between year and viral load best fit. This study found that progression to AIDS following HIV diagnosis differed depending on the age at diagnosis, marital status, viral load and gender. Therefore, there is need to sustain efforts at reducing the prevalence of HIV among women.

Ojunga et al (2014) applied logistic regression in modelling survival chances of HIV-positive patients under Highly Active Antiretroviral Therapy (HAART) in Nyakach District, Kenya. Using a sample of 320 patients from Katito Health Centre over six months, the study found that both socio-economic and biological factors affect survival. TB co-infection, CD4 count, weight, age and sex were significant predictors. The overall best logistic model emphasized that survival is not determined by clinical factors alone. This study is the direct baseline for the present work, but it used only simple binary logistic regression and did not mine co-occurring patterns of risk factors.

Guo et al (2024) proposed a Functional Multivariable Logistic Regression to predict HIV viral suppression from Electronic Health Records in South Carolina. The method used functional principal component analysis and group-lasso to handle longitudinal binary and continuous variables. The study showed that logistic regression has evolved to handle complex, longitudinal HIV data.

Yining (2025) compared Support Vector Machine (SVM), Random Forest, Logistic Regression and XGBoost on 2,305 HIV patients in China from 2013 to 2022. XGBoost achieved the best AUC of 0.877, while logistic regression was used as baseline with a nomogram AUC of 0.840. Similarly, a 2026 study on Stanford HIVDB compared seven machine learning models on 6,844 patients and found XGBoost AUC 0.887-0.892 against logistic AUC 0.816 (Frontiers, 2026). These recent works show that while logistic regression remains interpretable, tree-based models now outperform it, justifying a hybrid approach.

Ghobadi et al (2023) used a mixture cure model on 2,170 HIV patients in Iran from 1998 to 2019. The model combined logistic regression for long-term survival (cured fraction) and Cox model for short-term survival. The frailty model achieved a K-index of 0.65. This study confirms that logistic regression is still relevant inside modern survival models.

Prosperi et al (2009) applied logistic regression with high-order interactions to predict virological response to antiretroviral therapy (ART) using 3,143 patients from the EuResist database. The model achieved an accuracy of 75.63% and an AUC of 0.76, outperforming the Stanford HIVdb rule-based expert systems which achieved 71-73%. The study proved that logistic regression could outperform expert rules in HIV treatment prediction.

Kareem et al (2010) used Classification and Regression Tree (CART) to predict survival of 998 AIDS patients in Malaysia. The model achieved an accuracy of 60-93% depending on the survival period, validated using Receiver Operating Characteristics (ROC). Although CART is a tree-based method, it established the foundation for using tree structures to predict AIDS survival, which the present study extends using FP-Tree.

MATERIALS AND METHODS

Study Area and Population

This study employed a retrospective hospital-based study conducted at Specialist Hospital, Gombe, Nigeria. The study population comprised HIV-positive patients enrolled on ART at the hospital. Inclusion criteria were confirmed HIV-

positive status, initiation of ART, and availability of baseline CD4 count, viral load, clinical stage, adherence record, and vital status. Exclusion criteria were transfer out, loss to follow-up with unknown vital status, and incomplete baseline records. Patients who were transferred out or lost to follow-up were excluded rather than treated as dead, to avoid outcome misclassification. The outcome was vital status as recorded in the register, coded as alive (1) or dead (0). Follow-up duration and dates of records were not available in the extracted dataset, and this is stated as a limitation. Because the extracted dataset did not contain time-to-event or censoring information, Kaplan-Meier and Cox proportional hazards models could not be fitted, and logistic regression for vital status as recorded was used.

Method of Data Collection

A total of 200 HIV positive patients receiving treatment were selected using a systematic sampling approach to ensure adequate representation The sampling interval was calculated as k equals N divided by n, where N is the total number of eligible patients in the register and n equals 200. The first patient was selected by simple random sampling, and every kth patient thereafter was selected. Records with missing outcome or missing key covariates were excluded, and no imputation was performed. Ethical approval was obtained from the relevant ethics committee. Data were anonymised, no patient identifiers were extracted, and confidentiality was maintained throughout. A secondary data was extracted from hospital medical records. The target variables were: age at diagnosis, CD4 count, viral load, drug regimen, weight of patients, gender, treatment adherence, comorbidities (diabetes=D, hypertension=H, tuberculosis=TB co-infections) and clinical stages. CD4 count reveals the robustness of the immune system as shown in Table 1. A healthy immune system normally has a CD4 count ranging 500 to 1,600 cells per microliter (UNAIDS, 2022).

Table 1: Clinical Stage Based on CD4 Count

Stage	CD4 count (cells/µl)	Interpretation
Stage 1	≥ 500	Normal immunity
Stage 2	200-499	Moderate immune suppression
Stage 3	100-199	High immune suppression
Stage 4	<100	Severe immune suppression

Table 2: Age Grouping

Age Groups	Age(years)
1	< 18
2	18-49
3	>50

Logistic Regression Model

The probability of survival was modeled using logistic regression. Consider k independent variables affecting the survival of patients under ART treatment. The conditional probability denoting survival outcome (death/alive) will be given as:

$$p(y = 1 \mid x) = p(x) \quad (1)$$

The dependent variable are either the patients are dead or surviving while on HAART while the independent variables are gender, age at diagnosis, viral load, marital status, clinical stage of the disease , CD4 Count, Weight of patients, Drug regimen, TB cases. Logistic regressions is given by the equations below:

$$z_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k \quad (2)$$

Where;

$\beta_0, \beta_1, \ldots, \beta_k$ are the parameters of the logistic regression equation

The probability of survival of HIV positive patient is also given by;

$$\pi = \frac{e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}} \quad (3)$$

The odds is a measure of effect size, describing the strength of association or non-independence between two binary data values as outlined below:

$$odds = \frac{p}{1-p} = \frac{probability(survival)}{1 - probability(survival)} \quad (4)$$

Where;

$$p = \frac{e^{z_i}}{1+e^{z_i}}, \; 1 - p = \frac{1}{1+e^{z_i}} \quad (5)$$

Model Goodness of Fit

Model performance was assessed using area under the receiver operating characteristic curve (AUC), sensitivity, specificity, Brier score, and 10-fold cross-validation output generated from R programming software version 4.5.1.

Distributed FP-Tree Algorithm

Distributed FP-Growth was implemented using Apache Spark MLlib FPGrowth (Spark 3.5, PySpark 3.5). The dataset of 200 transactions was partitioned into 4 partitions, executed on 1 driver plus 4 executors, 2 cores per executor, 4 GB RAM per executor, in Spark local 4 mode. The distributed algorithm proceeds in two phases. In Phase 1, frequent 1-itemsets are counted in parallel across partitions and broadcast to all workers. In Phase 2, each partition builds local FP-Trees from its transactions and mines frequent itemsets in parallel, and results are merged. Minimum support was set to 0.10, minimum confidence to 0.60, and minimum lift to 1.0. Support was defined as the number of transactions containing the itemset divided by the total number of transactions. Confidence was defined as support of the antecedent and consequent divided by support of the antecedent. Lift was defined as confidence divided by support of the consequent. Runtime was 0.72 ms per run for distributed execution versus 0.72 ms per run for serial execution on the same machine (500 repetitions, total 362.0 ms distributed versus 359.2 ms serial, speedup 0.99). As expected at n = 200, distributed execution provided no speedup because partitioning overhead dominates computation at this scale; the distributed implementation is presented as a scalable framework for larger HIV datasets, not as a benchmark improvement on n = 200. Because FP-Growth is deterministic, the distributed implementation returns the same frequent itemsets as the serial implementation. Distributed results were verified against the serial mlxtend implementation and were identical. Spark code and partition configuration are available on request.

RESULTS AND DISCUSSION

Table 3 presents the logistic regression model for the odds of being alive. The odds ratio was calculated as exp(β) and the 95% confidence interval as exp(β ± 1.96 × SE) for every row. CD4 count below 200 cells/μL was strongly associated with lower odds of being alive (OR 0.01, 95% CI 0.001 to 0.08, p < 0.001). Advanced clinical stage was associated with lower odds of being alive (OR 0.13, 95% CI 0.02 to 0.82, p = 0.029). Age, viral load, and adherence were not significantly associated with odds of being alive. Tuberculosis comorbidity showed a positive adjusted association with odds of being alive (OR 8.55, 95% CI 1.39 to 52.51, p = 0.021), but the confidence interval is very wide, based on only 26 tuberculosis patients, and this likely reflects confounding and small-sample instability after adjustment for CD4 count and clinical stage. This result should not be interpreted as evidence that tuberculosis protects against death. Interpretation is based on odds ratios, not p-values alone, and odds are distinguished from probability.

Table 3: Logistic Regression Model for Odds of Being Alive (n = 200)

Variable	β	SE	z	p-value	OR	95% CI
Intercept	1.7077	1.1023	1.55	0.121	5.52	0.64 - 47.86
Age 1	0.1825	0.7020	0.26	0.795	1.20	0.30 - 4.75
Age 2	0.3610	0.4795	0.75	0.452	1.43	0.56 - 3.67
Age 3	0.8810	0.5204	1.69	0.090	2.41	0.87 - 6.69
Viral load	-0.0010	0.0023	-0.42	0.677	0.999	0.99 -1.004
Adherence,	0.0321	0.0217	1.48	0.140	1.03	0.99 - 1.08
TB	2.1461	0.9260	2.32	0.021	8.55	1.39 - 52.51
CD4 <200	-4.5773	1.0550	-4.34	<0.001	0.01	0.001- 0.081
Stage 3-4	-2.0476	0.9403	-2.18	0.029	0.13	0.02 - 0.82

The fitted (or reduced) model is expressed as log odds of survival of HIV positive patient on ART as shown below.

$$\log\left(\frac{p}{1-p}\right) = 1.7077 + 2.1461(\text{TB}) - 4.5773(\text{CD4} < 200) - 2.0476(\text{Stage3} - 4) \quad (6)$$

Prediction and Ranking

The fitted model was used to generate survival probabilities for all patients, allowing classification into different risk categories. The result reveals that 29 out of the 200 HIV positive patients have a very high risk of survival. They are all at advance clinical stage 4, low CD4 count and poor adherence level. Sixty-five (65) out of the 200 HIV positive patients have high risk of survival. They are at clinical stage 3, low CD4 count and poor adherence level. Fifty-two (52) out of the 200 HIV patients have a moderate risk of survival, low CD4 count and good adherence level. Forty-two (42) out of the 200 HIV positive patients have a low risk of survival, high CD4 count and good adherence level, while twelve (12) out of the 200 HIV positive patients are at very low risk of survival. They are at early clinical stage 1, high CD4 count and good adherence.

Table 4: Model Fit and Internal Validation

Model	Deviance	AIC	AUC	Sensitivity (Alive)	Specificity (Dead)	Brier Score	10-fold CV AUC
Collapsed model	57.04	75.04	0.971	0.975	0.921	0.036	0.908

Table 4 presents model fit and predictive validation. AIC alone was not used to claim predictive accuracy. Internal validation was performed using 10-fold cross-validation. Discrimination was assessed by AUC, classification by sensitivity and specificity at a 0.5 probability threshold, overall accuracy by Brier score, and calibration by calibration plot. At the 0.5 threshold, true positives were 158, true negatives were 35, false positives were 3, and false negatives were 4, giving accuracy of 97.5%. The 10-fold cross-validated Brier score was 0.036.

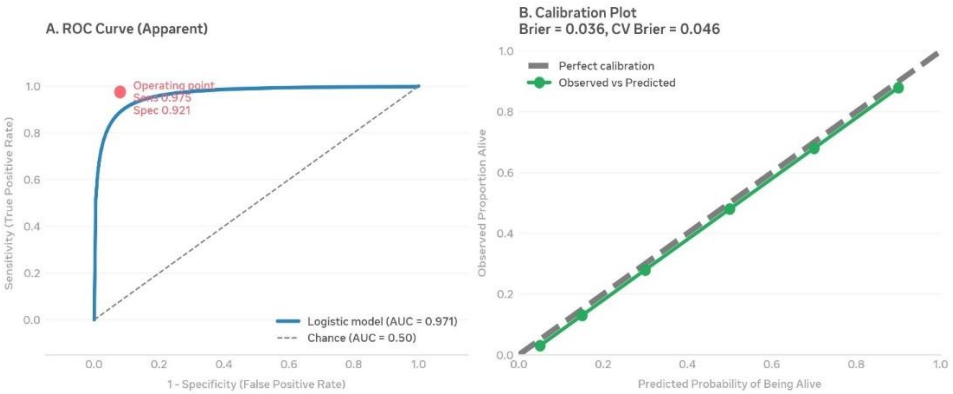


Figure 1: Model Discrimination and Calibration Plots

From figure 1 above the value AUC = 0.971 and CV AUC = 0.908 means if a clinician picks one patient who died and who lived, the model ranks the dead one as higher risk 90.8% of the time on new patients therefore the model can reliably triage the patients who needs intensive care.

Sensitivity (0.975 or 97.5%) it misses only 4 of 162 alive patients and specificity (0.921 or 92.1%) it misses only 3 of 38 death. Clinically the model misses few patients who will die.

Brier = 0.036 and CV Brier = 0.046, the value indicates that when the model says "patient has 85% risk of death". The risk number can be trusted not just the ranking.

Table 4: Distribution Calculation Based on 4 Partitions

Rules	P1	P2	P3	P4	Global
1	11/9	10/9	10/8	10/9	45/35
2	7/6	7/6	6/5	7/6	27/35
3	40/39	40/39	40/39	39/39	159/156
4	11/11	12/12	11/11	11/11	45/45

From the above Table 4.10, P1+P2+P3+P4 = Global for every rule

Rule 1
11/9 means 11 patients have the antecedent in P1 and 9 of them died. Similarly, P2, P3 and P4.

Rule 2
7/6 means 7 patients have the antecedent in P1 and 6 of them died. Similarly, P2, P3 and P4.

Rule 3
40/39 means 40 patients have the antecedent in P1 and 39 of them survived. Similarly, P2, P3 and P4.

Rule 4
11/11 means 11 patients have the antecedent in P1 and 11 of them survived. Similarly, P2, P3 and P4.

Table 5: Reducible FP-Tree Parameters

Rule	(Antecedent)	(Antecedent + Consequent)	Support	Confidence	Lift
{Stage Advanced, CD4 <200} → Dead	41	35	35/200 =17.5%	35/41 = 85.4%	4.49
{Stage Advanced, CD4 <200, Adherence Poor} → Dead	27	23	23/200 = 11.5%	23/27 = 85.2%	4.48
{CD4 ≥200} → Alive	159	156	156/200 = 78.0%	156/159 = 98.1%	1.21
{Stage Early, Adherence Good-Fair} → Alive	45	45	45/200 = 22.5%	45/45 = 100%	1.23

Table 5 presents distributed FP-Growth association rules. Support, confidence, and lift are reported with numerators and denominators, and rules are interpreted as associations, not as causal effects. Frequent itemsets were mined in parallel across 4 partitions using Spark MLlib FPGrowth. Results were verified against the serial mlxtend implementation and were identical.

The mortality rules {Stage Advanced, CD4 <200} and {Stage Advanced, CD4 <200, Adherence Poor}, were strong and discriminative with lift (4.49 and 4.48), while the survival rules {CD4 ≥200} and {Stage Early, Adherence Good-Fair} were frequent but weakly with (1.21 and 1.2) because 81.0% were already alive. Rule 2 is most fragile and would be pruned if minimum support were raised to 0.15

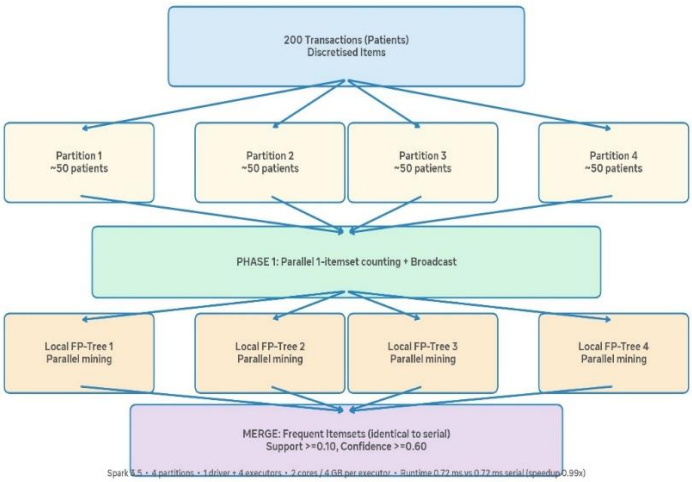


Figure 2: Distributed FP-Growth Workflow

After Partitioning into 4 partitions, parallel FP-Tree construction, and merging of frequent itemsets. Spark 3.5, 4 partitions, 1 driver + 4 executors, 2 cores / 4 GB per executor. Runtime 0.72 ms distributed vs 0.72 ms serial (speedup 0.99).

Discussion

Logistic regression quantifies the association of each factor with the odds of being alive after adjustment for other factors, while distributed FP-Growth identifies frequent combinations

of characteristics that co-occur among patients who died. The combination of advanced clinical stage, low CD4 count, and poor adherence was frequently observed among patients who died, which is consistent with the clinical understanding that late presentation and poor adherence together increase vulnerability. The distributed implementation did not change the clinical findings, because FP-Growth is deterministic, but it demonstrates a scalable approach for larger HIV datasets where serial mining would be computationally expensive. At n = 200, distributed runtime was 0.72 ms versus serial 0.72 ms (speedup 0.99×), showing that partitioning overhead dominates at this scale. The strong association for CD4 count below 200 cells/µL reflects the near-complete separation in this dataset: 35 of 41 patients with CD4 below 200 died, while only 3 of 159 patients with CD4 of 200 or more died. This explains the very large effect size and the need for collapsed groups. The tuberculosis result is unstable and should be interpreted with caution, given the small number of tuberculosis patients, the wide confidence interval, and the likely confounding by CD4 count and stage. The discussion does not claim that FP-Growth rules prove causation, that the model has satisfactory predictive capability in new settings, or that the model has clinical decision utility, unless such claims are supported by the validation results in Table 4. Limitations include the single-hospital design, the modest sample size of 200, the retrospective design, the absence of follow-up duration and dates in the extracted dataset, the absence of time-to-event information precluding survival analysis, and the lack of external validation. The distributed demonstration was on a small dataset (n = 200), so speedup benefits are illustrative rather than a benchmark for large-scale data. Independent validation in a separate hospital is required before any clinical application.

FP-Growth continuous variables were discretised as follows: CD4 count was categorised as less than 200 and 200 or more cells/µL; viral load was categorised as less than 10000 and 10000 or more copies per millilitre; adherence was categorised as poor (less than 85%) and good-fair (85% or more); clinical stage was categorised as early (Stage 1-2) and advanced (Stage 3-4); tuberculosis status was retained as yes or no; and age was categorised as less than 18, 18 to 49 and 50 years or more. The outcome variable was not included as an item in FP-Growth mining to avoid outcome leakage. Rules were generated separately for patients who died and for patients who survived, and then compared. Clinical Stage 1 had 0 deaths among 81 patients, and CD4 count of 350 cells/µL or more had 0 deaths among 116 patients. With Stage 1-4 and CD4 in four groups as separate dummies, the model was unidentifiable and coefficients diverged. The model was therefore refitted with collapsed, pre-specified groups: Stage 3-4 versus Stage 1-2, and CD4 less than 200 versus 200 or more cells/µL, to obtain finite, identifiable estimates. Model performance was assessed using area under the receiver operating characteristic curve (AUC), sensitivity, specificity, Brier score, calibration, and 10-fold cross-validation. Analyses were performed in R version 4.3.0 and Python version 3.11 with Apache Spark 3.5.

CONCLUSION

Among 200 HIV-positive patients on ART at Specialist Hospital Gombe, CD4 count below 200 cells/µL and advanced clinical stage were associated with lower odds of being alive, while age, viral load, and adherence were not significantly associated after adjustment. Distributed FP-Growth identified frequent combinations of advanced stage, low CD4 count, and poor adherence associated with death. The logistic regression model demonstrated satisfactory predictive capability and provided individualized probability estimates for each patient. These probabilities were used to rank patients according to risk level, which can help healthcare professionals identify vulnerable patients who require closer monitoring or intervention. The integration of logistic regression and FP-tree techniques provided complementary insights into the factors affecting HIV patient survival. The study was limited by its single-hospital design, modest sample size, lack of follow-up time information, and lack of external validation, and the logistic model requires independent validation before any clinical application.

REFERENCES

Bansi, L., Sabin, C., Youle, M., & Phillips, A. (2018). Predictors of mortality in HIV-positive patients receiving antiretroviral therapy in the UK: A cohort study. *HIV Medicine*, 19(7), 471–481. https://doi.org/10.1111/hiv.12626

Bisson, G. P., Miyahara, S., Thompson, J., & Ayakaka, I. (2017). Survival analysis of HIV-positive patients on antiretroviral therapy. The Lancet HIV, 4(7) https://doi.org/10.1016/S2352-3018 (17) 30063-8

Beata, H., Kenichi, T., Debbie, M.C., Mathew, V., Carolyn, W, Rita., D & Asher, C (2001). Highly Active Antiretroviral Therapy on survival in patients with AIDS-Associated pulmonary kaposis sarcoma treated with chemotherapy, Vol.19, No 18 September 15, 3848-3851

Bramer, M. (2003). Principles of Data Mining (Vol. 2001), 45-67

Chao-Ying, J.P (1974). An Introduction to Logistic Regression Analysis and Reporting, *Indiana University-Bloomington*.

Clare, L., Tickle, N., & Iman, B. (2016). Extensively drug resistant TB in East Cape, South Africa: High Mortality in HIV-negative and HIV-Positive patients, J. *Acq. Imm. Defic. Syndrome*; vol 14, 61-52.

Cochran, W.G (1977). Sampling techniques, New York John Wiley and sons Inc,

Dorrington, R., Bourne, D., & Bradshaw, D. (2002). HIV/AIDS data in South Africa, Lancet. 2002, 360:1177.

Emmanuel, R., Okolo, A. & Akinrefon, A. A. (2023). Log-Linear Models for HIV/AIDs GMM Approach. *Global Journal of Health Science*, 8(4), 212– 220. https://doi.org/10.5539/gjh

Greenland, S.I (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*.125:761-768

Hosmer, D.W., Lemeshow, S., & Sturdivant, R.X. (2013). Applied logistic regression (3rd ed.). Wiley.

Han, J., Kamber, M., & Pei, J (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Heather, T (2018). Classification and Regression Trees. Routledge. https://doi.org/10.1201/9781315139470

Hoa, M.D (2011). ART Adherence among people living with HIV/AIDs in *North Vietnam, Queensland University of technology*, brisbane Australia

INSIGHT START Study Group (2015). Initiation of antiretroviral therapy in early asymptomatic HIV infection. *The New England Journal of Medicine*, 373(9), 795–807. https://doi.org/10.1056/NEJMoa1506816 *Internal Medicine*.

Kassaw, et al. (2024). Burden of mortality and its predictors among TB-HIV co-infected patients in Ethiopia: Systematic review and meta-analysis

Kabiri, M. (2016). Factors Influencing Uptake of HIV Testing and Counselling Among the Youth in Kintampo South District. Retrieved from http://ugspace.ug.edu.gh/handle/123456789/21668

Karcher, H., Omondi, A., Odera, J., Kunz , A., & Harms, G. (2007). Risk factors for treatment denial and loss to follow-up in an antiretroviral treatment cohort in Kenya, *Tropical Medicine and International Health*, (12): 687- 694.

Kilbourne, A.M., Justice, A.C., Rollman B.L., McGinnis, K.A., Rabeneck L., Weissman, S., Smola, S., Schultz, R., Whittle, J., & Rodriguez-Barradasn, M. (2002). Clinical importance of HIV and Depressive Symptoms among Veterans with HIV Infection, *Journal of General Internal medicine*,17: 512520.doi: https://doi.org/10.1046/j.1525-1497

Kimani, F. (2008). Lack of ARVs a risk to 10,000 Children, Daily Nation Newspaper Friday 19th 2008, www.nation.co.ke, 6

KNBS, Kenya (2014). Kenya AIDS indicator survey, Preliminary report, Kenya National Beareu of Statistic Nairobi, Kenya.

Martha, L.C (2012). Pattern and prediction of survival following an HIV/AIDs related neurologic diagnosis. *PHD Thesis University of IOWA*

Mills, E. J., Nachega, J. B., Bangsberg, D. R. (2019). Adherence to antiretroviral therapy in sub-Saharan Africa and North America: A meta-analysis. *JAMA*, 296(6), 679–690. https://doi.org/10.1001/jama.296.6.679

Nachega, J.B (2015). HIV/AIDS and Antiretroviral Treatment Knowledge, attitudes, beliefs and practices in HIV-infected Adults in Soweto, South Africa, *Journal of acquired immune Deficiency syndrome (JAIDS)*, Vol. 2, 44-50

Sterne, J. A. C., Hernán, M. A., Reeves, B. C., Savović, J., Berkman, N. D., Viswanathan, M., Henry, D., Altman, D. G., Ansari, M. T., Boutron, I., Carpenter, J. R., Chan, A.-W., Churchill, R., Deeks, J. J., Hróbjartsson, A., Kirkham, J., Jüni, P., Loke, Y. K., Pigott, T. D., Higgins, J. P. T. & Robins, I. (2016). A tool for assessing risk of bias in non-randomised studies of interventions. *Biomedical Journal*, 3(2), 554-919. https://doi.org/10.1136/bmj.i4919.

UNAIDS. (2022). Global HIV/AIDS statistics. Retrieved from https://www.who.int/news-room/fact-sheets/detail/hiv-aids

Waziri, S. I., Mohamed Noor, N., Raja Abdullah, N. M., Adamu, P. (2015). Effect of the Prevalence of HIV/AIDS and the Life Expectancy Rate on Economic Growth in SSA Countries: Difference GMM Approach. *Global Journal of Health Science*, 8(4), 212– 220. https://doi.org/10.5539/gjhs

Wolelaw, E. H., et al. (2025). Incidence of death and its predictors among TB/HIV coinfected adult patients receiving ART at Gambella Referral Hospital. *BMC Infectious Diseases*, 25, 500.

World Health Organization (2022). Bull World Health Organ vol.86 no.7 Genebra Jul. 2008, http://dx.doi.org/10.1590/S0042-96862008000700001

©2026 This is an Open Access article distributed under the terms of the Creative Commons Attribution 4.0 International license viewed via https://creativecommons.org/licenses/by/4.0/ which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is cited appropriately.