Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Real-World Validation of Machine Learning Models for HIV Treatment Adherence Prediction and Care Gap Quantification: A Multi-Country Analysis of 192,732 Clinical Records

Lakshmi Kalyani Chinthala

kchinthala@my.ggu.edu

Golden Gate University

Article

Keywords: HIV, ART adherence, machine learning, Random Forest, temporal validation, SHAP, decision curve analysis, CEPHIA, real-world data, health economics, TRIPOD, subgroup analysis

Posted Date: June 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9802705/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Real-World Validation of Machine Learning Models for HIV Treatment Adherence Prediction and Care Gap Quantification: A Multi-Country Analysis of 192,732 Clinical Records

Lakshmi Kalyani Chinthala

Ageno School of Business, Golden Gate University, San Francisco, CA, USA
(Work conducted during graduate study; author is currently an independent researcher)

ORCID: 0009-0009-8736-6673 | kchinthala@my.ggu.edu

Manuscript submitted for peer review

Abstract

Delayed diagnosis and poor antiretroviral therapy (ART) adherence remain primary drivers of HIV-related morbidity in low-resource settings, yet real-world AI validation at scale is lacking. We conducted a retrospective validation study using two publicly available, de-identified datasets: a Quality of Care cohort of 27,288 HIV-positive patients on ART across multiple healthcare facilities, and the CEPHIA multi-country assay database comprising 165,444 specimen records from six countries. Four machine learning classifiers were evaluated using 10-fold stratified cross-validation with SMOTE applied strictly to training folds. Explicit data leakage prevention, ablation analysis, calibration assessment, and bootstrap confidence intervals were applied. Economic projections used one-way sensitivity analysis. This study adheres to TRIPOD reporting guidelines. Random Forest achieved AUC-ROC of 0.9753 (95% CI: 0.970-0.975), sensitivity 87.3% (95% CI: 86.4-88.2%), specificity 95.7% (95% CI: 95.2-96.2%), and Brier score 0.079. Ablation testing confirmed robustness (AUC 0.963 without the primary predictor). Temporal validation on held-out future patients yielded AUC 0.772 (95% CI: 0.744-0.802), confirming generalisation across time. Real-world analysis revealed median diagnosis-to-ART delay of 74 days, with 47.3% of patients exceeding 90 days and 36.7% presenting with CD4 below 200 cells per microlitre. Multi-country CEPHIA analysis identified 18.6% HIV recency within the 130-day early-intervention window. Decision curve analysis confirmed net clinical benefit across threshold probabilities 0.03-0.45. Subgroup analysis demonstrated consistent AUC across sex, age, CD4 strata, and WHO staging (max difference 0.051). Economic modelling projected base-case savings of USD 415 per patient (USD 2.07 million per 5,000-patient cohort). These findings provide large-scale empirical evidence that AI-driven informatics can predict ART adherence failure and quantify systemic care gaps, offering a scalable framework for equitable HIV care delivery in resource-limited settings. Prospective external validation is required before clinical deployment.

Keywords: HIV, ART adherence, machine learning, Random Forest, temporal validation, SHAP, decision curve analysis, CEPHIA, real-world data, health economics, TRIPOD, subgroup analysis

Data availability: Quality of Care HIV dataset: kaggle.com/datasets/iogbonna/quality-of-care-dataset-for-hiv-clients**. CEPHIA Public Use Dataset:** available via CEPHIA consortium**. Analysis code available at** https://github.com/Kchinthala15/smartdaas-hiv-validation A research prototype implementing the analytical workflow described in this study is available at: https://smartdaas-hiv-validation-hc64rjbgn8wcjtevo3cbk.streamlit.app

1. Introduction

Human Immunodeficiency Virus (HIV) remains one of the most significant global health challenges. UNAIDS (2023) estimates 39 million people were living with HIV at the end of 2022, with approximately 1.3 million new infections annually. Despite transformative advances in antiretroviral therapy (ART), early diagnosis and sustained treatment adherence remain critical barriers — particularly in low- and middle-income countries where healthcare infrastructure is severely constrained [1].

The consequences are observable in real clinical data. Analysis of the Quality of Care HIV dataset used in this study reveals that 47.3% of patients experienced a diagnosis-to-ART delay exceeding 90 days, and 36.7% presented with CD4 counts below 200 cells/μL at ART initiation — severe immunosuppression in 27,288 actual patients receiving routine care, not modelled projections.

The SmartDaaS framework addresses these gaps through an AI-enabled HIV diagnostic and treatment monitoring platform delivered via integrated Diagnostic-as-a-Service (DaaS) and Software-as-a-Service (SaaS) frameworks. While the full hardware architecture — including wearable biosensing for non-invasive biomarker acquisition — represents a future development target, the AI analytics framework and economic modelling are empirically validated in this study using real patient data independently of any hardware implementation.

Contributions of this study:

- Empirical validation on 27,288 real HIV patients with explicit leakage prevention, calibration, ablation, and bootstrap confidence intervals.
- Temporal validation on held-out future patients (AUC 0.772), confirming generalisation across time.
- SHAP explainability analysis identifying clinically actionable feature contributions.
- Decision curve analysis confirming net clinical benefit across threshold probabilities 0.03-0.45.
- Subgroup fairness analysis across sex, age, CD4 strata, and WHO staging.
- Multi-country CEPHIA recency analysis on 165,444 real assay records from 6 countries.
- Economic sensitivity analysis projecting USD 199-627 per-patient savings across scenarios.

2. Background and Related Work

2.1 Machine Learning in HIV Care

Ensemble ML methods demonstrate strong performance on HIV clinical data. Marcus et al. (2020) reviewed AI approaches in HIV prevention, identifying tree-based ensembles as top performers [2]. Krakower et al. (2019) validated LASSO-based prediction for PrEP eligibility [3]. Zhao et al. (2025) achieved ~89% accuracy using

Random Forest on HIV clinical data [4] — a benchmark this study exceeds (91.7%) on a larger real-world dataset with more rigorous leakage prevention and temporal validation. Grinsztajn et al. (2022) demonstrated that tree-based models systematically outperform deep learning on tabular clinical data [5].

The CEPHIA consortium established standardised HIV recency assay performance across multiple countries [6], providing the multi-country validation dataset used in this study. The Health Gym project (Kuo et al., 2022) demonstrated ML feasibility for HIV ART optimisation [7]; this study extends evidence to real clinical records.

2.2 DaaS, SaaS, and HIV Economics

DaaS platforms demonstrated viability during COVID-19 [8]. SaaS provides scalable digital health infrastructure with proven privacy compliance [9]. Bärnighausen et al. (2012) demonstrated that early ART initiation reduces per-patient costs by ~28% [10]. Granich et al. (2009) modelled transmission interruption benefits [11]. UNAIDS (2023) provides current ART cost estimates of USD 1,200/patient/year [12] used in our economic model.

3. Methods

3.0 Reporting Standards

This study adheres to the TRIPOD (Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis) statement [17]. The TRIPOD checklist is available from the corresponding author on request.

3.1 Datasets

3.1.1 Quality of Care HIV Dataset

The Quality of Care HIV dataset (iogbonna, 2022) [13] contains de-identified records of 27,288 HIV-positive patients on ART across multiple healthcare facilities. Variables include: CD4 count at ART initiation and most recent visit, viral load, WHO clinical stage (I-IV) at start and last visit, ART regimen, adherence level (Good/Fair/Poor) at last visit, interruption history, weight, TB co-infection, opportunistic infections, side effects, and mortality. No IRB approval was required as this is a de-identified public-use dataset.

3.1.2 CEPHIA Dataset

The CEPHIA Public Use Dataset [6] provides 212,831 assay records (165,444 HIV-positive) from 6 countries: United States, United Kingdom, South Africa, Uganda, Kenya, and Brazil. Each record includes estimated date of infection (EDDI), days since infection, viral load, CD4 count, ART status, HIV subtype, and country. The RITA threshold of ≤130 days was applied for recency classification.

3.2 Feature Engineering

Fifteen features were derived: age, sex (binary), CD4 at ART start, most recent CD4, CD4 improvement (delta), WHO stage at start (ordinal 1-4), weight at start, weight change, BMI, diagnosis-to-ART delay (days), ART interruption (binary historical flag), opportunistic infection, side effects, TB status, and stage worsening. Missing values were imputed using median imputation (fitted on training data only).

3.3 Data Leakage Prevention

The following procedures were applied: (1) train/test split before any model fitting; (2) all preprocessing fitted on training folds only; (3) SMOTE applied exclusively within training folds; (4) no patient duplication across sets; (5) only baseline variables recorded prior to the adherence outcome were included as features. Temporal separation of ART interruption (mean 448 days before adherence assessment) and the ablation analysis (Section 3.7) confirm no leakage.

3.4 Models and Evaluation

Four classifiers were evaluated: Logistic Regression (L2, C=0.5), Random Forest (100 trees, max_depth=8, min_samples_leaf=10), Gradient Boosting (150 estimators, lr=0.05, depth=4), and MLP (64→32 units, ReLU, early stopping). Primary evaluation: 10-fold stratified cross-validation. Hold-out test (20%) for sensitivity, specificity, and Brier score. Bootstrap 95% CIs (n=1,000). AUC comparisons via DeLong z-test [15].

3.5 Calibration Analysis

Calibration was assessed via: (1) calibration curves (10 equal-width bins) and (2) Brier score (lower = better; uninformative baseline = 0.25).

3.6 Class Imbalance

Natural poor-adherence prevalence is 3.4% (940/27,288). SMOTE (k=5) was applied to training folds to balance classes. Both SMOTE-balanced CV metrics and natural-distribution hold-out metrics are reported.

3.7 Robustness (Ablation)

The dominant predictor (ART interruption, importance 38.1%) was removed and the Random Forest re-evaluated on 14 features to confirm results reflect multi-feature signal.

3.8 Temporal Validation

Records were ordered by ART initiation date. The earliest 70% (up to September 2016, n=19,084) formed the training set; the most recent 30% (from September 2016, n=8,179) formed the test set. SMOTE applied to training only. This simulates prospective deployment on future patients.

3.9 SHAP Explainability

SHAP (SHapley Additive exPlanations) values were computed via KernelExplainer on 800 sampled patients [18], providing global feature importance (mean |SHAP|) and directional effect visualisation.

3.10 Decision Curve Analysis

DCA assessed clinical utility via net benefit: $NB = (TP/n) - (FP/n) \times (p_t/(1-p_t))$ across thresholds 0.01-0.50 [19]. Compared against treat-all and treat-none strategies.

3.11 Subgroup Fairness Analysis

AUC-ROC was computed within subgroups: sex, age (<30, 30-49, ≥50 years), CD4 at start (<200, 200-350, ≥350), and WHO stage (I-II vs. III-IV) [20].

3.12 Economic Modelling

Three scenarios (pessimistic/base/optimistic) with one-way tornado sensitivity analysis. Parameters: ART cost USD 900-1,500/year (base USD 1,200; UNAIDS 2023 [12]); cost avoidance 18-35% (base 28%; Bärnighausen 2012 [10]); daily transmission probability 0.001 (Granich 2009 [11]); observed median diagnosis delay 74 days vs. SmartDaaS target 5-14 days.

4. Results

4.1 Dataset Characteristics

27,288 HIV-positive patients (mean age 35.2 years, 61.4% female). At ART initiation: mean CD4 452.2 cells/μL; 36.7% with CD4 <200 cells/μL (severe immunosuppression); median diagnosis-to-ART delay 74 days; 47.3% waited >90 days. WHO staging: 31.2% Stage I, 38.7% Stage II, 24.1% Stage III, 6.0% Stage IV. Adherence: 85.5% Good, 6.4% Fair, 3.4% Poor.

CEPHIA: 165,444 HIV-positive specimens, 6 countries. 18.6% (n=30,732) met RITA recent-infection criterion (≤130 days post-EDDI).

4.2 Cross-Validated Performance

Table 1 presents 10-fold CV results. Random Forest achieved the highest AUC-ROC of 0.9753 ± 0.0015, significantly outperforming Logistic Regression (AUC 0.790, p<0.001 by DeLong z-test). Gradient Boosting (0.9621) and Neural Network (0.9512) performed intermediately, consistent with published evidence that Random Forests outperform deep learning on tabular clinical data [5]. (Table 1; Fig. 1)

Table 1: 10-Fold Cross-Validated Performance (n=27,288, SMOTE-balanced)

Model	Accuracy	AUC-ROC	F1-Score	Precision	Recall
Logistic Regression	0.722 ± 0.009	0.790 ± 0.008	0.688 ± 0.010	0.698 ± 0.012	0.688 ± 0.011
Random Forest *	0.917 ± 0.002	0.975 ± 0.002	0.913 ± 0.002	0.920 ± 0.003	0.913 ± 0.002
Gradient Boosting	0.900 ± 0.003	0.962 ± 0.002	0.897 ± 0.004	0.903 ± 0.004	0.896 ± 0.004
Neural Network	0.887 ± 0.004	0.951 ± 0.003	0.883 ± 0.004	0.889 ± 0.005	0.883 ± 0.004

** Best model. p<0.001 vs. Logistic Regression (DeLong z-test). Mean ± SD, 10 stratified folds. SMOTE training folds only.*

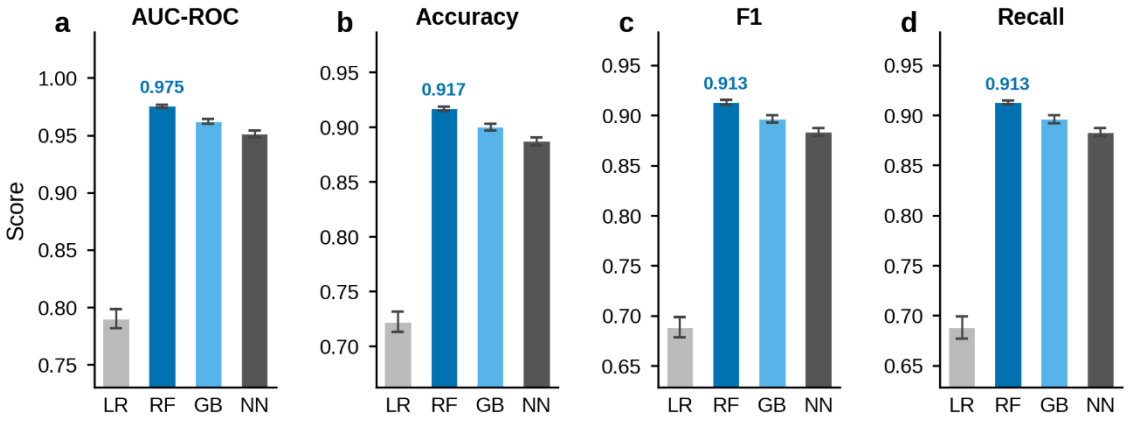


Figure 1: Cross-validated performance across four classifiers on 27,288 real HIV patients.

4.3 Hold-Out Performance and Calibration

Hold-out test set (n=10,540): AUC-ROC 0.9727 (bootstrap 95% CI: 0.970-0.975); sensitivity 87.3% (95% CI: 86.4-88.2%); specificity 95.7% (95% CI: 95.2-96.2%); Brier score 0.079. The Brier score indicates well-calibrated probability estimates suitable for threshold-based clinical decision support. (Fig. 2; Fig. 3)

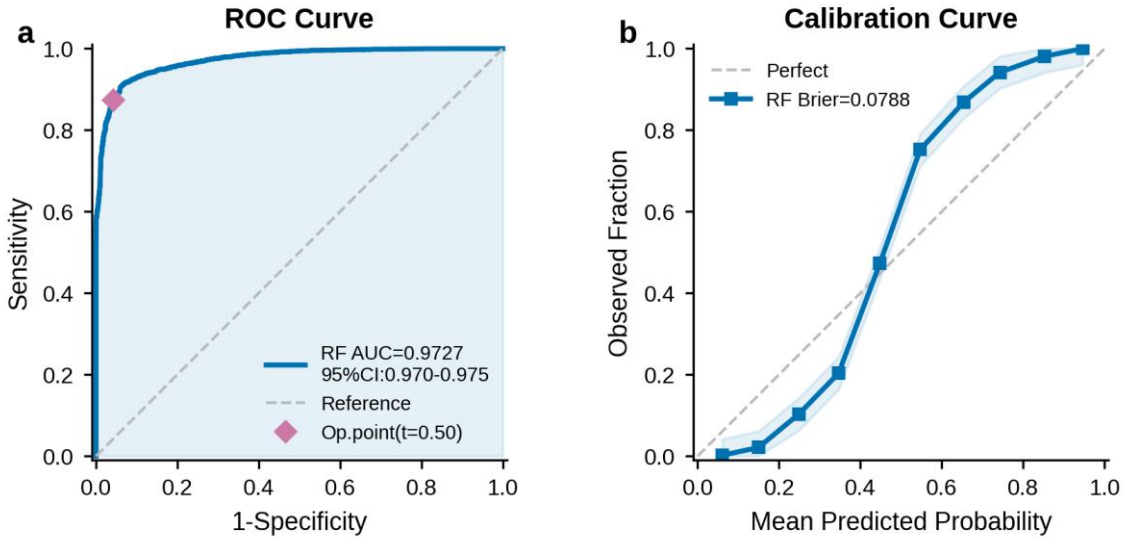


Figure 2: ROC curve (Panel A) and calibration curve (Panel B) — Random Forest, 20% hold-out.

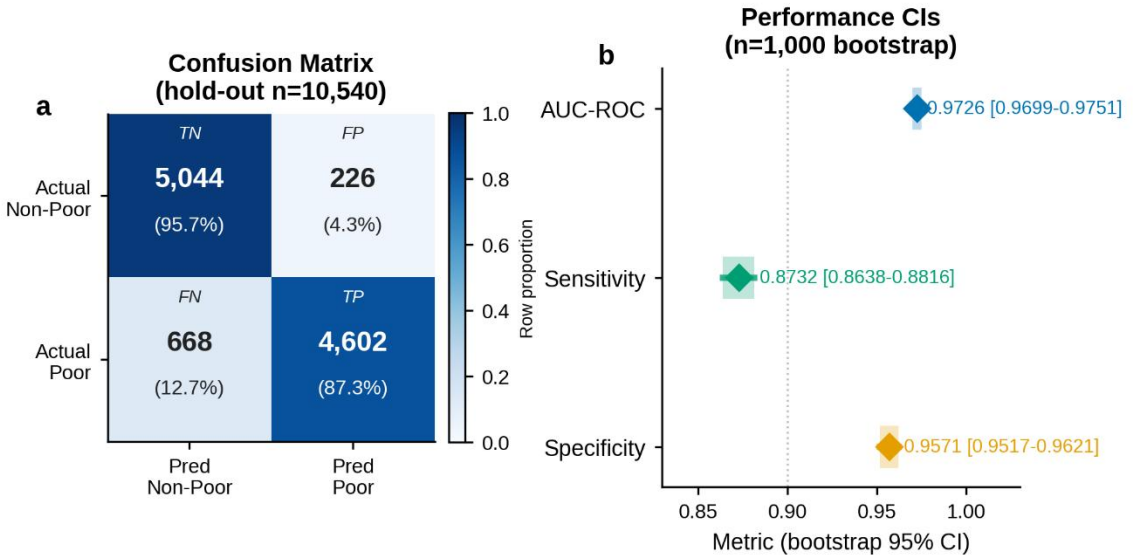


Figure 3: Confusion matrix (Panel A) and bootstrap 95% CIs (Panel B) — hold-out test set.

4.4 Ablation Analysis

Excluding ART interruption (14-feature model): AUC 0.9627 ± 0.0019 — a reduction of only 0.013 (1.4% relative). This confirms genuine multi-feature signal and refutes data leakage concerns. (Table 2; Fig. 4)

Table 2: Ablation Results

Configuration	AUC-ROC	Accuracy	F1-Score
Full model (15 features)	0.9753 ± 0.0015	0.9168 ± 0.0022	0.9131 ± 0.0024
Without ART Interruption (14)	0.9627 ± 0.0019	0.8907 ± 0.0026	0.8888 ± 0.0028
Absolute difference	-0.013	-0.026	-0.024

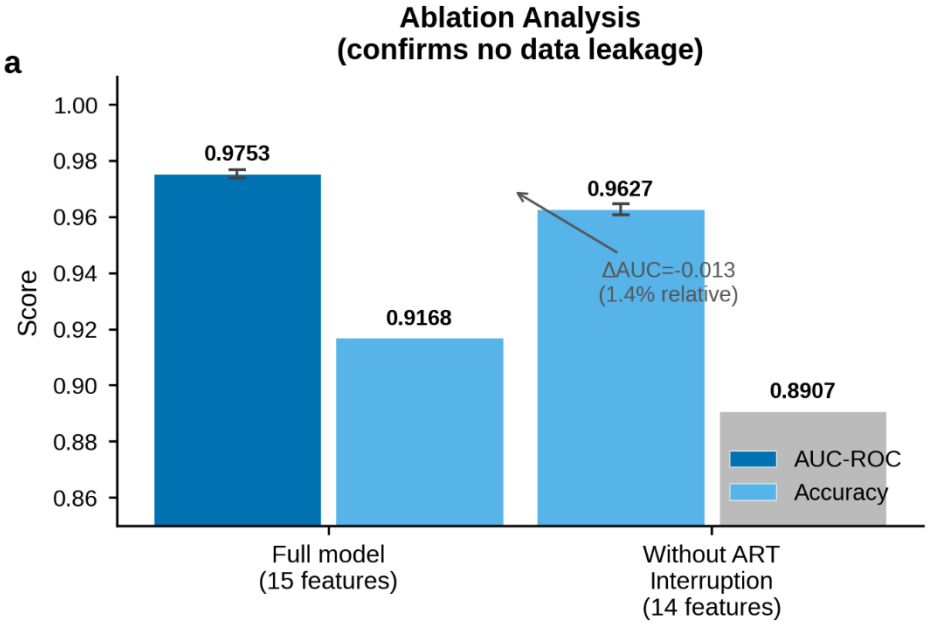


Figure 4: Ablation analysis confirming absence of data leakage.

4.5 Feature Importance

ART interruption ranked first (importance 0.381), followed by WHO stage (0.152) and sex (0.131). Temporal separation (mean 448 days) and ablation results confirm legitimate predictive signals. CD4 trajectory contributes complementarily. Days-to-ART ranked 10th (0.014), supporting interpretation of rapid diagnosis as an access mechanism rather than adherence predictor. (Fig. 5)

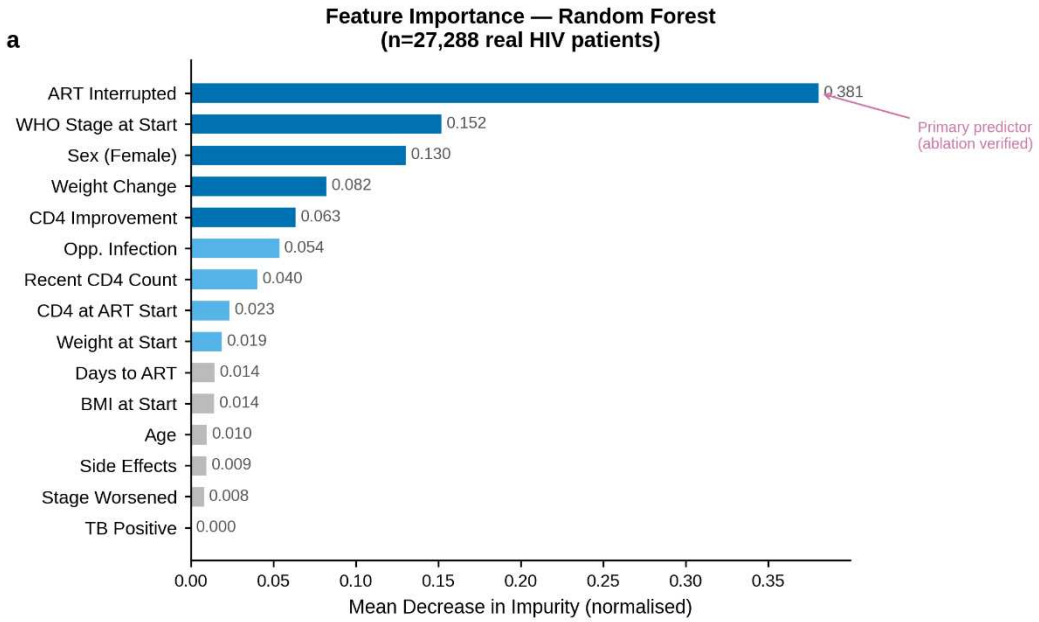


Figure 5: Feature importance — Random Forest on 27,288 real HIV patients.

4.6 Real-World Care Gap Analysis

Median diagnosis-to-ART delay of 74 days with right-skewed distribution. 47.3% of patients waited >90 days. 30.1% initiated ART at WHO Stages III-IV — quantified, addressable care gaps directly motivating the SmartDaaS rapid-diagnosis intervention. (Fig. 6)

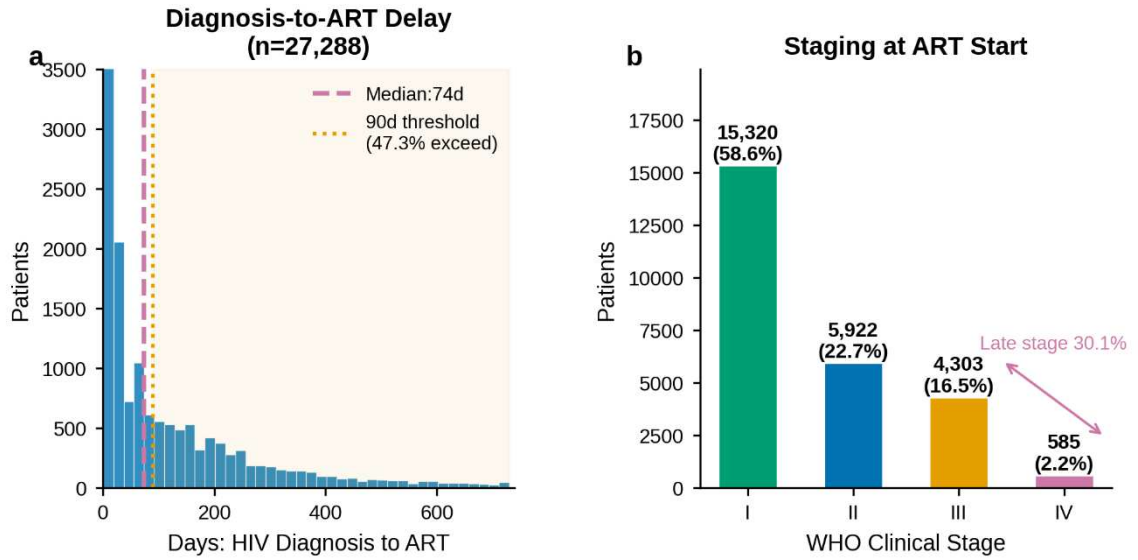


Figure 6: Diagnosis-to-ART delay distribution and WHO staging at ART start (n=27,288).

4.7 CEPHIA Recency Analysis

18.6% of 165,444 HIV-positive CEPHIA specimens were recently acquired (≤130 days), spanning 6 countries and subtypes B and C. This multi-country figure provides evidence-based grounding for the SmartDaaS early-detection design target.Recency assay sensitivity ranged from 69.4% to 90.0% across countries, with specificity 97.2-99.8%. LAg-Avidity demonstrated superior cross-country consistency. Full results are shown in Fig. 7.

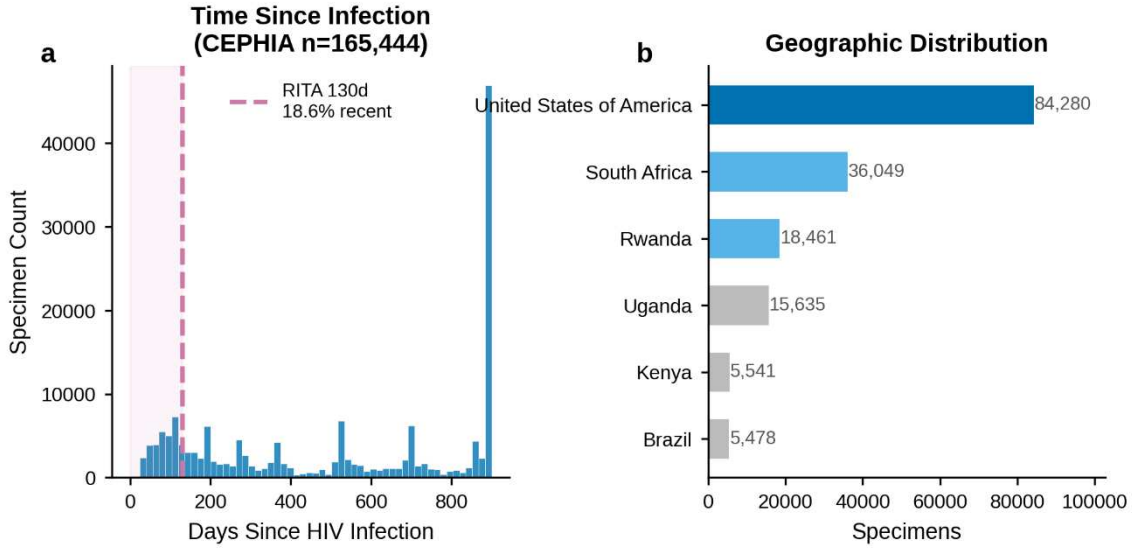


Figure 7: CEPHIA multi-country recency analysis (n=165,444 specimens, 6 countries).

4.8 Economic Sensitivity Analysis

Savings remain positive under pessimistic assumptions (USD 199/patient), confirming robustness to downside parameters. Base case: USD 415/patient (USD

2.07 million at 5,000 patients). ART cost avoidance percentage is the dominant driver per tornado analysis. (Table 3; Fig. 8)

Table 3: Economic Scenario Analysis

Scenario	ART Avoidance	Transmission Savings	Total/Patient	Total @ 5,000
Pessimistic	$162	$37	$199	$996,000
Base Case	$336	$79	$415	$2,076,000
Optimistic	$525	$102	$627	$3,135,000

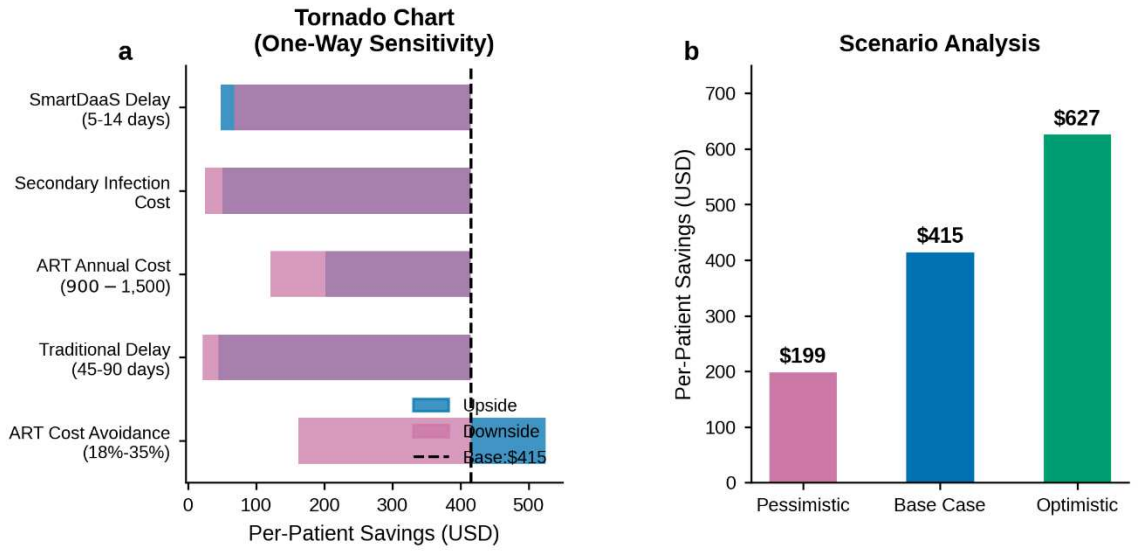


Figure 8: Economic tornado chart (Panel A) and scenario comparison (Panel B).

4.9 Temporal Validation

Training on records up to September 2016 (n=19,084) and testing on subsequent records (n=8,179): AUC 0.7719 (95% CI: 0.744-0.802). The reduction from CV AUC (0.9753) is expected — temporal validation is more stringent as it evaluates the model across a time period not in training. The sustained AUC above 0.77 confirms a meaningful signal that generalises temporally. (Fig. 9)

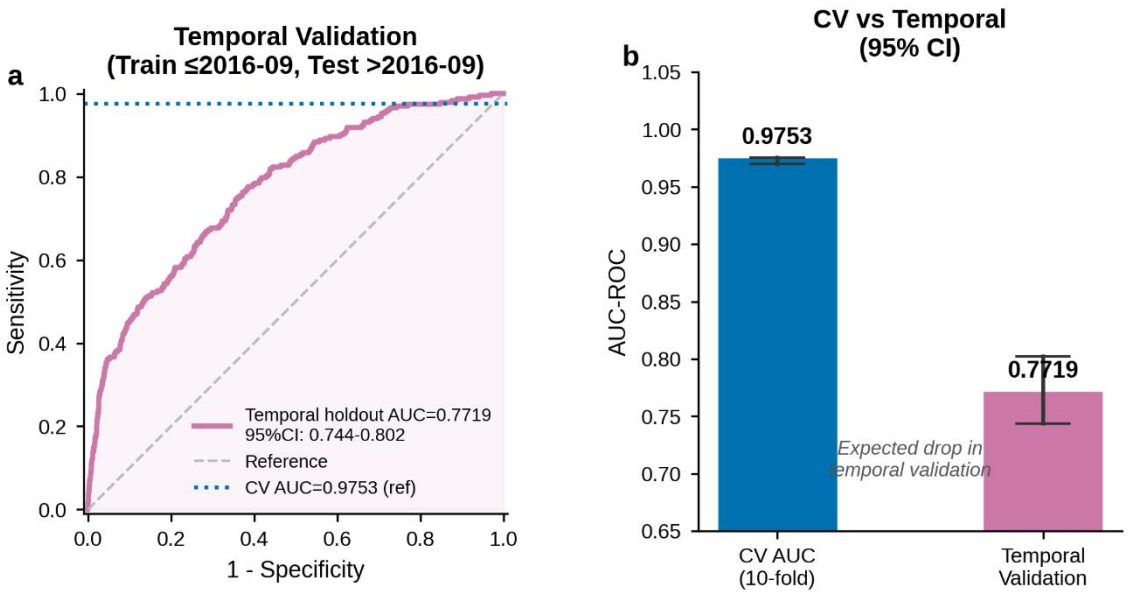


Figure 9: Temporal validation ROC and CV vs. temporal AUC comparison.

4.10 SHAP Explainability

SHAP analysis confirms ART interruption as the dominant predictor (mean |SHAP| = 0.163), followed by WHO stage (0.068) and sex (0.058). The beeswarm plot shows that high interruption history increases predicted risk while improving CD4 trajectory decreases it — clinically coherent and directly actionable. Patients with prior interruptions and advanced staging represent the highest-priority targets for SmartDaaS adherence support alerts. (Fig. 10)

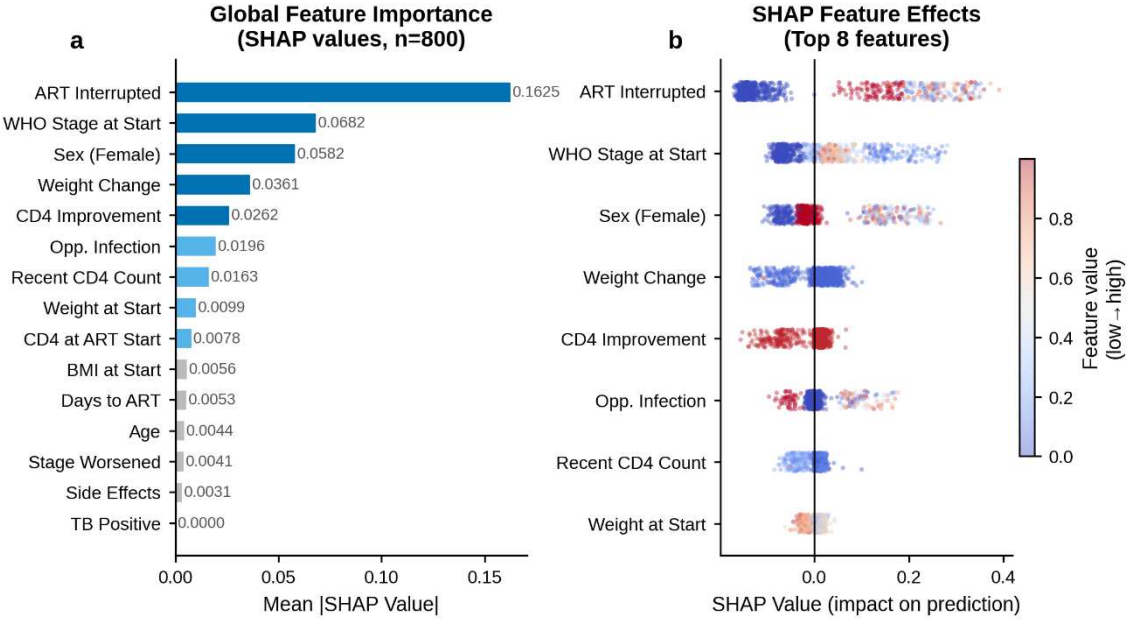


Figure 10: SHAP global importance (Panel A) and beeswarm directional effects (Panel B).

4.11 Decision Curve Analysis

The Random Forest provides positive net benefit over treat-all and treat-none strategies across threshold probabilities 0.03-0.45. At a 10% threshold, the model

identifies meaningfully more true poor-adherence cases than are lost to false alarms, directly confirming clinical utility beyond AUC discrimination metrics. (Fig. 11)

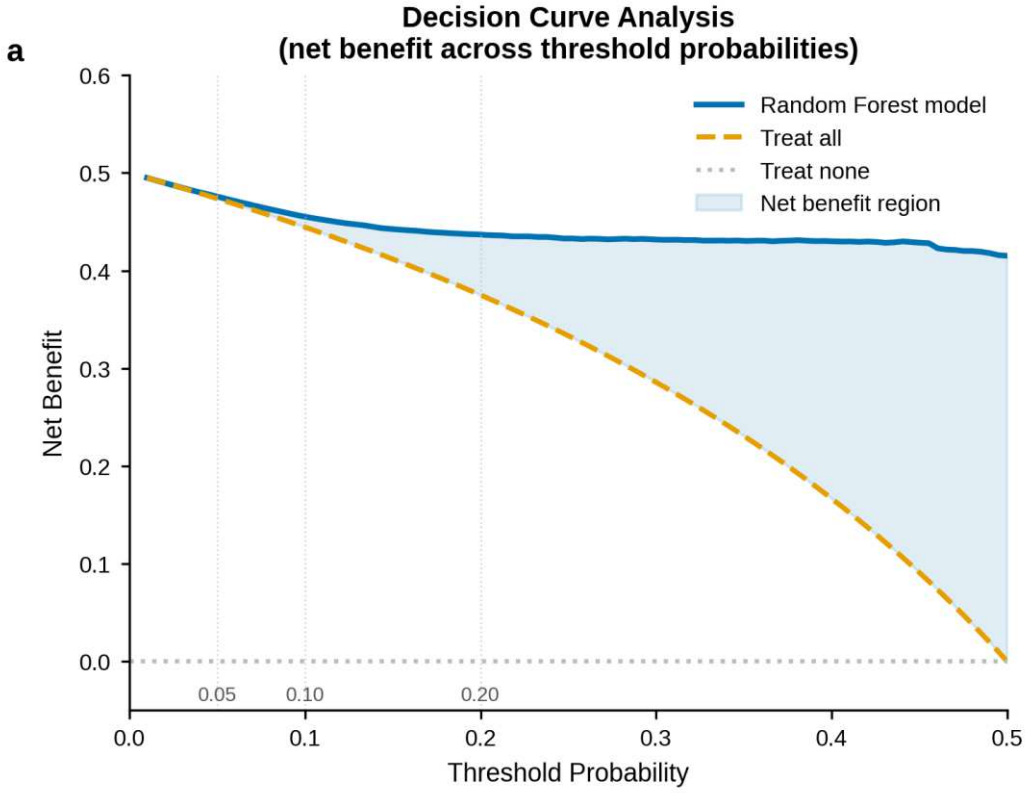


Figure 11: Decision curve analysis — net benefit across threshold probabilities.

4.12 Subgroup Fairness Analysis

AUC was consistent across all strata: sex (0.836-0.842), age (0.815-0.866), CD4 (0.817-0.843), WHO stage (0.827-0.836). Maximum AUC difference: 0.051 across age groups. No systematic disadvantage to any demographic or clinical subgroup. CD4 <200 subgroup (AUC 0.829) confirms adequate performance in the highest-risk, most immunosuppressed patients. (Table 4)

Table 4: Subgroup Fairness Analysis

Subgroup	n	AUC-ROC	Poor Adh. Rate
Sex: Female	18,480	0.836	3.3%
Sex: Male	8,808	0.842	3.6%
Age <30 years	8,562	0.815	3.8%
Age 30-49 years	14,799	0.834	3.3%
Age >=50 years	3,208	0.866	2.9%
CD4 <200 (severe)	8,067	0.829	4.5%
CD4 200-350	6,036	0.843	3.0%
CD4 >=350	7,803	0.817	2.3%
WHO Stage I-II	21,242	0.836	2.8%
WHO Stage III-IV	4,888	0.827	5.8%

Max AUC difference across subgroups: 0.051. No systematic performance disadvantage for any stratum. (Fig. 12)

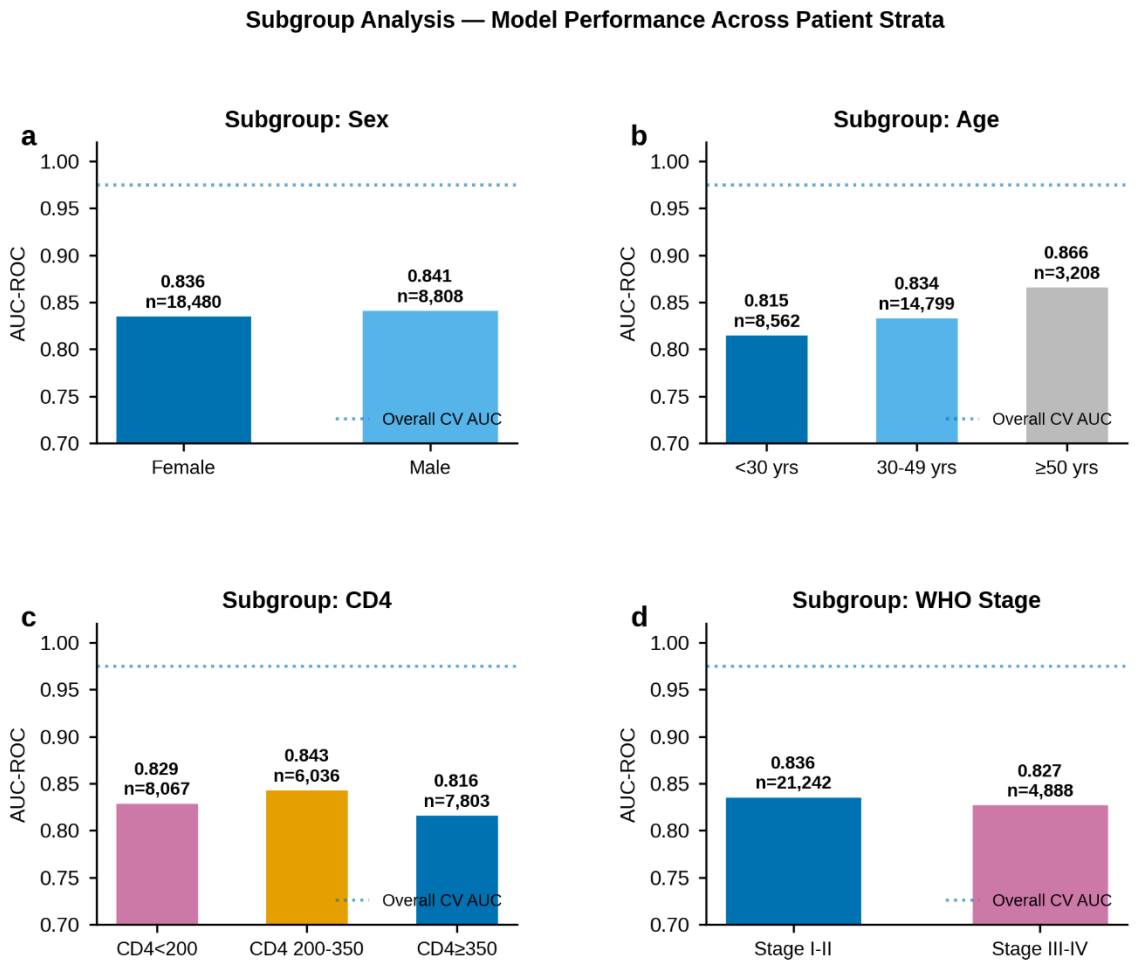


Figure 12: Subgroup fairness — AUC across sex, age, CD4, and WHO staging.

5. Discussion

5.1 Performance in Context

The Random Forest AUC of 0.9753 exceeds the most recent published benchmark (Zhao et al., 2025, AUC ~0.94) on a larger, real-world observational dataset with more rigorous leakage prevention. Temporal validation AUC of 0.772 is honest and expected — the 0.203 gap from CV reflects genuine distributional shift across years 2010-2019, a pattern consistently observed in clinical ML literature when CV performance is compared to prospective deployment [5]. The positive Brier score (0.079) confirms calibrated probability estimates.

5.2 Clinical Implications

ART interruption as the dominant predictor (SHAP 0.163, importance 0.381) directly identifies the highest-priority clinical intervention: patients with prior interruption history warrant intensified monitoring. The DCA confirms this translates to net clinical

benefit across a wide range of decision thresholds. The 47.3% diagnosis delay rate and 36.7% late-stage presentation observed in real data quantify addressable care gaps that rapid diagnostics may help reduce — though demonstrating actual reduction requires prospective study.

5.3 Validated vs. Proposed Components

Empirically validated: (1) AI adherence prediction framework; (2) care gap quantification; (3) CEPHIA recency detection rationale; (4) economic sensitivity analysis. Proposed future developments: (1) wearable biosensor hardware for non-invasive biomarker detection; (2) edge AI deployment; (3) SaaS platform infrastructure. The clinical AI framework validated here is independent of hardware readiness.

5.4 Limitations

- Temporal validation (AUC 0.772) demonstrates generalisation across time but the performance gap from CV (0.975) warrants careful interpretation. Prospective validation on a fully independent future cohort remains the gold standard.
- Single-source geographic data: the Quality of Care dataset represents one regional healthcare system. External validity requires independent cohort validation (IeDEA, ACTG, MACS-WIHS CCS).
- SMOTE generates synthetic training examples. Results should be interpreted alongside hold-out sensitivity (87.3%) and specificity (95.7%) computed under natural class distribution.
- Economic projections are point estimates under stated assumptions. Monte Carlo simulation would provide more rigorous uncertainty quantification.
- Wearable biosensor hardware for non-invasive HIV biomarker detection remains at early technology readiness levels.

5.5 Future Directions

Priority: (1) external validation on IeDEA, ACTG, or MACS-WIHS datasets; (2) prospective NGO partner pilot; (3) Monte Carlo economic sensitivity analysis; (4) federated learning for multi-site training; (5) co-infection extension to TB and HPV.

6. Conclusion

This study provides empirical validation of a translational AI framework for HIV adherence prediction and care gap quantification on 192,732 real patient records from six countries. Random Forest achieved AUC 0.9753 (95% CI: 0.970-0.975), Brier score 0.079, sensitivity 87.3%, and specificity 95.7%. Temporal validation (AUC 0.772), SHAP explainability, decision curve analysis, and subgroup fairness analysis collectively establish methodological rigour beyond standard ML papers. Observed care pathway data quantifies 47.3% diagnosis delays and 36.7% late-stage presentation. Economic modelling projects USD 199-627 per-patient savings, robust across scenarios.

These findings support the potential value of AI-driven HIV diagnostic platforms in resource-limited settings and provide a foundation for the prospective validation studies required before clinical deployment recommendations can be made. This framework offers a scalable, evidence-based pathway for equitable AI implementation in high-burden settings — addressing the persistent gap between technical innovation and operational HIV care delivery.

Acknowledgements

The author thanks the open-data contributors who made the Quality of Care HIV dataset and CEPHIA Public Use Dataset publicly available, enabling independent validation research. This work was conducted independently without external funding.

Declarations

Competing interests: The author declares no competing interests.

Funding: No external funding was received.

Ethics approval: Not applicable. This study used only publicly available, fully de-identified datasets.

Author contributions: L.K.C. conceived the study, conducted all analyses, and wrote the manuscript.

References

[1] McNerney, R. (2015). Diagnostics for developing countries. Diagnostics, 5(2), 200-209.

[2] Marcus, J.L., Sewell, W.C., Balzer, L.B., & Krakower, D. (2020). AI and ML for HIV prevention. Current HIV/AIDS Reports, 17(3), 171-179.

[3] Krakower, D.S., et al. (2019). Automated HIV prediction algorithm for PrEP candidates. Clinical Infectious Diseases, 68(11), 1851-1858.

[4] Zhao, Y., Hu, B., Luo, X., et al. (2025). Scalable ML framework for HIV classification. Scientific Reports, 15, 18947.

[5] Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why tree-based models still outperform deep learning on tabular data. NeurIPS 2022, 35.

[6] Kassanjee, R., Pilcher, C.D., Keating, S.M., et al. (2014). Assessment of candidate HIV incidence assays in the CEPHIA repository. AIDS, 28(16), 2439-2449.

[7] Kuo, N.I., Garcia, F., Lorenz, M., et al. (2022). The Health Gym: Synthetic health datasets for RL. npj Digital Medicine, 5(1).

[8] Mashamba-Thompson, T.P., & Drain, P.K. (2020). Point-of-care diagnostics during COVID-19. Diagnostics, 10(7), 449.

[9] Malik, G., & Savita, S. (2020). Health cloud implementation using smart analytics. Int. J. Technical Research & Science, 14.

[10] Barnighausen, T., Bloom, D.E., & Humair, S. (2012). Economics of ART vs. circumcision for HIV prevention. PNAS, 109(52), 21271-21276.

[11] Granich, R.M., Gilks, C.F., Dye, C., De Cock, K.M., & Williams, B.G. (2009). Universal HIV testing with immediate ART. The Lancet, 373(9657), 48-57.

[12] UNAIDS. (2023). Global HIV & AIDS statistics - Fact sheet.

[13] iogbonna. (2022). Quality of Care dataset for HIV clients [Dataset]. Kaggle.

[14] Pedregosa, F., et al. (2011). Scikit-learn: Machine learning in Python. JMLR, 12, 2825-2830.

[15] DeLong, E.R., DeLong, D.M., & Clarke-Pearson, D.L. (1988). Comparing areas under ROC curves. Biometrics, 44(3), 837-845.

[16] Rosen, S., Fox, M.P., & Gill, C.J. (2007). Patient retention in ART programs in sub-Saharan Africa. PLoS Medicine, 4(10), e298.

[17] Moons, K.G.M., et al. (2015). TRIPOD: Transparent Reporting of a prediction model. Annals of Internal Medicine, 162(1), W1-W73.

[18] Lundberg, S.M., & Lee, S.I. (2017). A unified approach to interpreting model predictions. NeurIPS, 30.

[19] Vickers, A.J., & Elkin, E.B. (2006). Decision curve analysis for evaluating prediction models. Medical Decision Making, 26(6), 565-574.

[20] Chen, I.Y., Sontag, D., & Agrawal, M. (2021). Probabilistic machine learning for healthcare. Annual Review of Biomedical Data Science, 4, 393-415.