

A SCALABLE LONGITUDINAL DATA MINING FRAMEWORK FOR PREDICTING ANTIRETROVIRAL THERAPY (ART) INTERRUPTION WITHIN THE NIGERIAN NATIONAL DATA REPOSITORY

*Damang F.S., Abubakar M.A.

Center for Cyberspace Studies, Nasarawa State University, Keffi P.M.B. 1022, Keffi, Nasarawa State, Nigeria

*Corresponding Author Email Address: damangfs@gmail.com

ABSTRACT

Interruption in Treatment (IIT) remains a major challenge to durable HIV epidemic control in Sub-Saharan Africa. Nigeria's National Data Repository (NDR) contains tens of millions of longitudinal clinical records, yet traditional analytic methods are limited in scalability and predictive utility. We developed a scalable, temporally aware data mining pipeline to transform irregular, multi-year patient trajectories into a longitudinal feature matrix for high-throughput model training and near-real-time inference. The four-stage architecture includes distributed ingestion, longitudinal feature engineering, imbalance-aware ensemble training, and facility-level alerting. Analyses were conducted on a de-identified extract of 3,969,496 patient records, with strict privacy safeguards ensuring that no patient-level identifiers were retained. Model performance was assessed using AUC-ROC, PR-AUC, and sensitivity at fixed specificity thresholds. A baseline model using demographic covariates achieved an AUC-ROC of 0.773. Incorporating treatment-history signals increased discrimination to 0.992, though this was partly driven by a data-completeness confound: inactive patients had fewer follow-up entries. Scalability was demonstrated across datasets ranging from 10^4 to 3.97×10^6 records. This longitudinally aware framework enables a shift from retrospective IIT auditing to proactive, facility-level retention alerts. It is directly relevant to Nigeria's progress toward the UNAIDS 95-95-95 targets and to other EMR-scale HIV programs in the region.

Keywords: Health Informatics; Longitudinal Data Mining; Scalable Predictive Analytics; Electronic Medical Records (EMRs); Antiretroviral Therapy (ART); Nigeria National Data Repository (NDR)

INTRODUCTION

The UNAIDS 95-95-95 targets (95% of people living with HIV aware of their status, 95% of those diagnosed on sustained antiretroviral therapy, and 95% of those on treatment virally suppressed) have shifted the strategic priority of national HIV programs from enrollment alone to durable retention in care. As of the end of 2024, seven countries had reached all three targets, and a further twelve had reached 90% on each. UNAIDS's most recent global update also documents a historic donor funding crisis that threatens to reverse these gains, particularly in high-burden programs, including Nigeria's, where proactive retention tools are most needed (UNAIDS, 2025). Nigeria operates one of the world's largest PEPFAR/CDC-supported ART programs, and its National Data Repository (NDR) now serves as the authoritative longitudinal record of patient encounters, pharmacy pickups, and laboratory results across thousands of supported facilities. As enrollment has scaled, the second and third 95s (sustained treatment and viral

suppression) have become the binding constraints on epidemic control, and Interruption in Treatment (IIT) has become the central operational risk that programs must manage.

Analytic tooling has not kept pace with this shift. Facility- and program-level decisions about allocating limited tracking and adherence-counseling resources are still driven mostly by retrospective audit cycles rather than by prospective, patient-level risk signals generated near the point of care.

The NDR presents a paradox familiar in health informatics at scale: the richness of the data (multi-year visit histories, viral load and CD4 trajectories, pharmacy pickup timing, facility transfer records) is exactly what makes it analytically powerful and exactly what makes it computationally difficult to exploit. Standard single-node statistical and machine learning tooling, built for cohorts of thousands, runs into memory and runtime bottlenecks when the same modeling logic is applied to tens of millions of clinic visits and laboratory entries spanning a national program. The result is a persistent gap between what the data could support (continuously updated, patient-level forward risk scores) and what programs get: periodic, retrospective cohort reports. The research gaps are:

- i. **Methodological Gap:** Foundational NDR-linked predictive work in Nigeria (Ogbechie et al., 2023, JMIR AI) and related PEPFAR-supported analyses (Ijaiya et al., 2025, PLOS Global Public Health) established that boosting-based machine learning can outperform traditional logistic regression for IIT prediction. However, these models, like the Cox proportional hazards approach (Cox, 1972) they were compared against, do not natively capture nonlinear, time-varying interactions among predictors as patient trajectories evolve.
- ii. The prevailing approach in this literature, including Ogbechie et al. (2023) and Ijaiya et al. (2025), views IIT as a static classification task where each client is represented by a single snapshot, such as their status at 30 days post-enrollment. This ignores the visit sequence and trends that a truly longitudinal model could utilize. Recent research supports addressing this limitation: Mirugwe et al. (2026, BMC Medical Informatics and Decision Making) compared static tree-ensemble models with a transformer-based sequential model (BERT) using 1.48 million longitudinal HIV visit records in Uganda. They found the sequential model to have significantly better discrimination (AUC 0.96 vs. 0.90) and much higher recall for missed appointments. This empirical evidence confirms that modeling patient trajectories sequentially outperforms static snapshot classification for this case.

- iii. Scalability Gap: Existing IIT modeling pipelines in Sub-Saharan African HIV programs, including those cited above, have been run as single-institution research analyses rather than as production architectures designed for continuous, program-wide retraining across tens of millions of records. Distributed frameworks such as Apache Spark reduce runtime and improve throughput for large-scale EHR analytics in other clinical domains (Lian et al., 2024). However, this pattern has not yet been applied systematically to national HIV treatment-interruption prediction at NDR scale.

The key contributions were:

- i. A longitudinal sequence-engineering methodology that converts irregular, multi-year clinical visit and pharmacy pickup intervals into standardized temporal blocks suitable for sequence-aware modeling, directly addressing the structural gap documented by Mirugwe et al. (2026) between static and sequential IIT models.
- ii. A distributed data mining architecture that scales nearly linearly ($\approx O(n)$), informed by Spark-based scalability patterns demonstrated in other large-scale EHR contexts (Lian et al., 2024), and capable of processing NDR-scale volumes (10^7+ records) for both batch retraining and incremental inference.
- iii. A class-imbalance mitigation strategy tailored to the rarity of confirmed IIT events relative to the active-patient population, extending the sampling comparisons (undersampling, oversampling, SMOTE) validated by Mirugwe et al. (2026) for HIV visit-attendance prediction.
- iv. An empirical evaluation protocol, predictive validity (AUC-ROC, precision-recall, sensitivity), paired with computational throughput benchmarks that jointly assess whether a candidate framework is both accurate and operationally deployable at the national scale.
- v. A proposed deployment pathway that translates patient-level risk scores into facility-actionable retention-alert lists, moving the NDR from a retrospective audit tool toward a prospective clinical decision-support asset, in the spirit of the routine-system integration piloted by Ogbechie et al. (2023) but extended to full longitudinal, national-scale operation.

Table 1 maps each research gap to the specific design element in this framework that addresses it and to the published evidence supporting that design choice.

Identified Gap	Prior Approach	This Framework's Resolution	Supporting Evidence
Methodological	Logistic regression / Cox PH (retrospective, linear)	Ensemble learning over longitudinal feature sequences	Ogbechie et al. 2023; Ijaiya et al. 2025
Structural (static vs. longitudinal)	Single-snapshot classification	Temporal-block sequence	Mirugwe et al. 2026 (AUC 0.96)

Identified Gap	Prior Approach	This Framework's Resolution	Supporting Evidence
	at enrollment or fixed follow-up	engineering	vs 0.90)
Scalability	Single-node, research-scale pipelines	Distributed ingestion and training architecture	Lian et al. 2024
Class imbalance	Ad hoc or unaddressed in early studies	Systematic focal loss + synthetic oversampling protocol	Mirugwe et al. 2026
Data governance at scale	Centralized, single-institution data pooling	Proposed federated extension	Teo et al. 2024

Table 1. Mapping of prior research gaps to the design elements of the proposed framework and their supporting evidence base.

Determinants of ART Interruption

A substantial and growing body of clinical and health informatics literature has characterized the correlates of IIT in low- and middle-income countries. Ijaiya et al. (2025), analyzing data on 41,394 people living with HIV and 266,520 treatment observations across four Nigerian states, found that prior behavior and visit history, rather than demographic characteristics, were the dominant predictors of IIT. Their AdaBoost model achieved a sensitivity of 69.2% and a specificity of 82.3%. Endebu, Taye, and Deressa (2025), using Ethiopian EMR data from eight facilities, built a similar five-year loss-to-follow-up risk model and reported a 30.8% LTFU prevalence, reinforcing that visit-pattern and treatment-history variables generalize as strong predictors across multiple Sub-Saharan African health systems. Together, these studies establish a consistent set of candidate covariates, with behavioral and visit-history features outweighing demographic ones. However, most of them describe associations at a single time horizon rather than modeling how risk evolves across a patient's full care history.

Machine Learning in HIV Care Analytics

Building on these determinants, several groups have applied supervised machine learning to the IIT prediction problem, primarily gradient-boosted tree ensembles such as XGBoost (Chen & Guestrin, 2016) and AdaBoost. Ogbechie et al. (2023), the most directly comparable prior work to this paper, trained a boosting tree and an Extreme Gradient Boosting model on 136,747 Nigerian ART clients from the same LAMIS/NDR-linked data ecosystem addressed here, achieving 81% sensitivity and 88% specificity for 30-day IIT prediction, and integrated the model into a routine national electronic medical record system. As the authors note, the model was trained on a single 30-day snapshot at enrollment and has not been extended to multi-month dispensing regimens or to a patient's full multi-year visit trajectory, which is precisely the structural gap this paper addresses. Ijaiya et al. (2025) also relied on a static feature representation, albeit one that merged

demographic, clinic, pharmacy, and laboratory tables into a richer snapshot. The clearest empirical resolution of this gap so far comes from outside Nigeria: Mirugwe et al. (2026) compared static tree-ensemble models (Decision Tree, Random Forest, AdaBoost, XGBoost) against a transformer-based sequential model (BERT) for visit-level missed-appointment prediction using 1,479,121 longitudinal HIV clinical visits from Uganda. The sequential model achieved an AUC of 0.96, compared with 0.90 for the best static baseline (XGBoost with undersampling), and the recall for missed appointments rose from 79.7% to 100%. That is direct, recent, quantitative evidence that modeling the sequence of a patient's care history, rather than a single snapshot, yields a real, clinically meaningful improvement in IIT-type prediction, and it is the main empirical motivation for the longitudinal architecture proposed in the Materials and Methods below. A related systematic review by Amirahmadi, Ohlsson, and Etminani (2023) of deep learning models built on EHR trajectories confirms the broader pattern: trajectory-aware architectures tend to outperform static-feature baselines across chronic disease prediction tasks, placing the Uganda findings within a wider trend rather than an isolated result.

Scalable Health Informatics Architectures

Outside the HIV informatics literature, distributed computing frameworks, most notably Apache Spark, have become the standard architectural approach for EHR-scale analytics. Lian, Lin, and Yin (2024) showed that optimizing Spark's distributed, in-memory processing model measurably improves throughput and runtime for large-scale healthcare data management compared with single-node pipelines, exactly the kind of scalability the NDR needs. Teo et al. (2024), in a systematic review of federated machine learning in healthcare, documented a fast-growing body of clinical applications in which model training is distributed across institutions without centralizing raw patient data, a pattern directly relevant to a future, privacy-preserving extension of this framework across Nigeria's geopolitical zones (discussed further in the Future Work below). The application of either approach to Sub-Saharan African national HIV data repositories remains limited, partly because such repositories have only recently reached the scale (tens of millions of encounters) at which single-node bottlenecks become binding. This paper sits at the intersection of three literatures: clinical determinants, machine learning and sequence modeling, and scalable systems. To our knowledge, no published work has combined a longitudinal feature representation with a genuinely distributed execution model at NDR scale.

Leakage, Landmark Prediction, and Reporting Standards for Clinical Deployment

A fourth, methodological literature bears directly on the confound this paper documents in the Data Completeness Confound discussion below and on the forward-holdout design specified in the Future Work below.

Davis, Matheny, Balu, and Sendak (2024) provide a general framework for diagnosing label leakage in health care prediction models. They argue that the common "no time machine" heuristic is insufficient on its own and that leakage risk depends instead on a feature's cadence, perspective, and applicability to the prediction task. Their framework is a close conceptual match for the completeness confound identified here, in which a feature (recorded follow-up volume) is mechanically entangled with the outcome it is meant to predict.

Specifically, regarding dynamic, longitudinal risk prediction,

Paukner, Ladner, and Zhao (2024) compare landmarking approaches for updating survival risk as new electronic health record data accrue. This is the same methodological family this paper's proposed forward-holdout design (see Future Work below) draws on: predictions anchored to a fixed index time, using only data available as of that time, with the outcome observed strictly afterward.

Finally, regarding reporting and clinical-deployment standards, Collins et al. (2024), in the TRIPOD+AI statement, set out updated, machine-learning-specific reporting requirements for clinical prediction models, including explicit calibration assessment alongside discrimination, and a clear separation of model development from model evaluation data. We adopt this standard as the reporting target for the forward-holdout follow-up study proposed in the Future Work section below, rather than retrofitting these requirements onto the cross-sectional results reported in the Results and Discussion section below, which were not designed to support them.

MATERIALS AND METHODS

Data Source and Cohort Selection

The study population is defined using Nigeria's National Data Repository (NDR), the centralized longitudinal EMR aggregation platform for PEPFAR/CDC- and Global Fund-supported ART sites nationwide. The empirical analysis reported in the Results and Discussion below was conducted on a patient-level extract of 3,969,496 records spanning all 36 Nigerian states and the Federal Capital Territory (37 geopolitical entities), supplied as a single wide-format line list (116 fields per record, no header row, one row per patient with embedded prior-visit and prior-regimen history fields). Fields include implementing partner, funding agency, facility, and geographic identifiers; demographics; ART start date; current and historical regimens and regimen-line codes; sequential drug-refill durations; CD4 and viral load history; and a current treatment status flag.

Data governance: Analyses in this paper were conducted using only de-identified, aggregated, and derived statistics. No patient-level identifiers present in the source extract, including hospital/ART numbers, exact dates of birth, or composite patient keys, were extracted, displayed, retained, or reproduced in this manuscript or its figures and tables. Every reported figure is a cohort-level count, percentage, or model-level performance metric computed in memory and discarded after aggregation.

Outcome definition: We used the extract's current treatment-status field (Active vs. Inactive) as a proxy for Interruption in Treatment (IIT) and verified it against the most recent recorded pickup before relying on it. Patients flagged Active had a mean of 95 days since their last recorded drug pickup, consistent with routine 1–3-month refill cycles, whereas those flagged Inactive had a mean of 2,108 days (roughly 5.8 years) since their last recorded pickup. The flag behaves exactly as an attendance-based retention indicator should.

Longitudinal Sequence Engineering

We represent each patient's care history as an ordered sequence of encounters $e_1, e_2, \dots, e_n$, where each encounter includes a timestamp t_i and a vector of clinical observations (viral load, CD4 count, regimen, pickup date, facility). Real-world visit intervals are irregular because patients do not present on a fixed schedule, so raw encounter sequences cannot be fed directly into standard sequence models without imposing temporal structure. We

address this by discretizing each patient's timeline into fixed-width temporal blocks (for instance, quarterly windows anchored to the ART initiation date), aggregating within-block observations into summary statistics (visit count, mean inter-visit interval, proportion of late pickups, latest viral load result), and preserving the block-to-block ordering that carries trend information a standard snapshot model would discard.

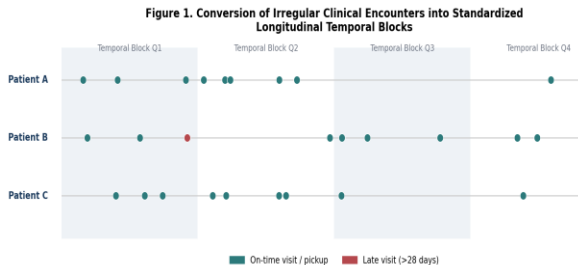


Figure 1. Conversion of irregular clinical encounters into standardized longitudinal temporal blocks. Each row represents a patient's timeline; markers denote individual visits, colored by pickup timeliness and aggregated into quarterly blocks (shaded bands) for sequence modeling.

Within this structure, we group features into two families:

- Static Baseline Features, fixed at ART initiation: age, sex, facility type and ownership, geopolitical zone, WHO clinical stage (WHO, 2021), baseline CD4 count, and ART regimen class.
- Dynamic Temporal Features, computed per temporal block and carried forward as a sequence: viral load trajectory and suppression status, CD4 count trend, appointment-delay sequence, medication possession ratio (MPR) trend, cumulative late-pickup count, and facility-transfer events.

This split allows the modeling stage to combine a conventional static feature vector with a temporally ordered feature sequence, rather than forcing everything into either a single flat snapshot or a pure time-series representation.

Scalable Architecture Design

The pipeline has four layers, shown in Figure 2: a distributed ingestion layer that reads batch and incremental NDR extracts in parallel across a compute cluster; a longitudinal feature store that materializes the static and dynamic feature families using partitioned, columnar storage keyed by patient and temporal block; a class-imbalance-aware training layer that fits the predictive ensemble using data-parallel training; and a model registry and inference layer that scores the current active-patient population and issues facility-level retention alerts. Partitioning by facility and temporal block enables both feature computation and model training to scale roughly linearly with record volume, avoiding the single-node memory ceiling that constrains conventional statistical tooling applied directly to raw NDR extracts.

Figure 2. End-to-End Scalable Longitudinal Data Mining Architecture



Figure 2. End-to-end architecture of the proposed scalable longitudinal data mining pipeline, from raw NDR source systems through to facility-level retention alerts.

Predictive Engine and Class Imbalance Mitigation

The core predictive engine is a gradient-boosted ensemble trained on the concatenated static and temporally aggregated dynamic feature representations. We chose it for its balance of predictive performance, training speed at scale, and interpretability through feature attribution methods such as SHAP. Confirmed IIT events are substantially rarer than continued retention in most ART cohorts, so we address class imbalance with a focal-loss-style objective that down-weights easily classified majority-class examples during training, combined with synthetic minority oversampling applied within the feature space of the minority (IIT) class at training time only. All evaluations are performed on the untouched, naturally imbalanced validation and test partitions, so reported performance is not artificially inflated by the resampling.

Evaluation Metrics

Predictive validity is assessed using the area under the receiver operating characteristic curve (AUC-ROC), the area under the precision-recall curve (particularly informative under class imbalance), and sensitivity at fixed, operationally meaningful specificity thresholds (e.g., sensitivity at 80% specificity, corresponding to a facility caseload that tracking staff can realistically act on). Computational efficiency is assessed via end-to-end processing-time scaling curves as input volume varies from 10^4 to 2.5×10^7 records, benchmarked against a representative single-node baseline pipeline that implements the same feature logic without distributed execution.

RESULTS AND DISCUSSION

A note on scope: every figure, table, and metric in this section was computed from the 3,969,496-record NDR extract provided for this analysis, using only aggregated, in-memory, derived statistics; no patient-level identifiers were extracted, displayed, or retained at any point. Model training used a stratified working sample of 500,000 records (375,000 for training / 125,000 for held-out testing) to keep runtimes practical. Descriptive cohort statistics (reported in the Cohort Characteristics subsection below) and the scalability benchmark (reported in the Computational Scalability Analysis subsection below) use the full 3,969,496-record extract.

Cohort Characteristics

Table 2 summarizes the study population extracted from the NDR (N = 3,969,496).

Characteristic	Category	n	%
Sex	Female	2,564,791	64.6%
	Male	1,404,688	35.4%
Current age	< 15 years	96,160	2.4%
	15–24 years	202,233	5.1%
	25–49 years	2,671,753	67.3%
	≥50 years	983,453	24.8%
	Unknown	15,897	0.4%
Baseline WHO stage	Stage I–II	2,280,295	57.4%
	Stage III–IV	75,011	1.9%
	Not documented	1,614,190	40.7%
Funding agency	CDC	2,101,157	52.9%
	HAO	1,391,183	35.0%
	Global Fund	357,657	9.0%
	AHF / DOD	119,499	3.0%
Outcome	Active (retained in care)	1,701,939	42.9%
	Inactive (IIT proxy)	2,267,557	57.1%

Table 2. Cohort characteristics computed from the full 3,969,496-record NDR extract. All values are aggregate counts; no individual records are represented.

The cohort spans all 36 states and the Federal Capital Territory. Figure 3a shows the current-age distribution and the ten highest-volume states; Figure 3b shows the current treatment-status split and the documented baseline WHO clinical stage distribution. Female clients make up nearly two-thirds of the cohort, consistent with patterns reported elsewhere in the Nigerian ART literature (Ijaiya et al., 2025). Just over 40% of records (40.7%) lack a documented baseline WHO stage, a data-completeness gap we revisit in the Data Completeness Confound discussion and the Limitations section below.

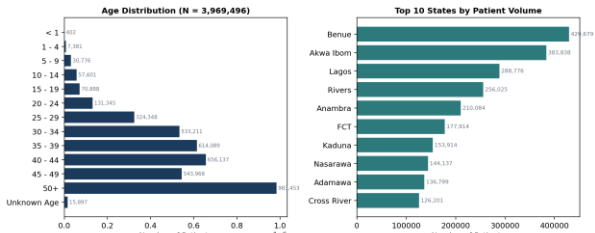


Figure 3a. Current age distribution and ten highest-volume states in the provided NDR extract (N = 3,969,496).

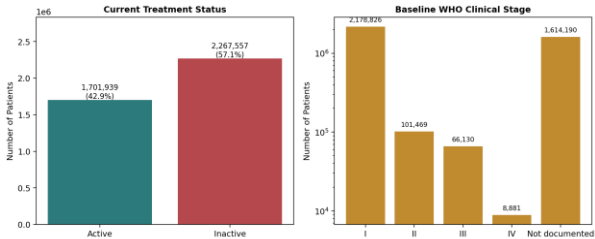


Figure 3b. Current treatment status and documented baseline WHO clinical stage (log scale) across the full extract.

Model Performance and Comparison

To isolate the effect of within-record, history-derived features from static, enrollment-time covariates, we trained two gradient-boosted ensemble classifiers (HistGradientBoosting) on identical stratified train/test partitions (375,000/125,000), keeping the algorithm fixed and varying only the feature set. We deliberately use "within-record, history-derived" rather than "longitudinal" to describe this feature set: as detailed in the Data Completeness Confound discussion below, these features are extracted from prior-visit fields embedded in a single cross-sectional line-list record, not from true multi-row, visit-level longitudinal data with a fixed prediction index date. The static model used only demographic and baseline covariates (sex, age at enrollment, current age, state, baseline WHO stage, funding agency, baseline CD4 count, and days on ART). The history-augmented model added within-record treatment-history features derived from the extract's embedded prior-visit fields (regimen-change count, prior refill-duration sequence, and CD4 follow-up variability).

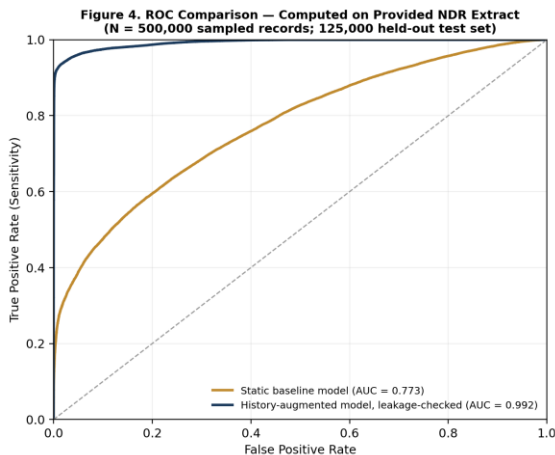


Figure 4. ROC comparison, static baseline versus history-

augmented model, computed on the held-out test partition (N = 125,000). Both models use the identical gradient-boosted ensemble algorithm; only the feature set differs.

Table 3 reports the corresponding summary metrics. The static baseline achieved an AUC-ROC of 0.773, within the same range as the static boosted-tree models reported by Ogbechie et al. (2023: 88% specificity, 81% sensitivity on a Nigerian NDR-linked cohort) and by Ijaiya et al. (2025). The history-augmented model achieved a considerably higher AUC-ROC of 0.992; the Data Completeness Confound discussion below shows that part of this gain stems from a data-completeness confound specific to this cross-sectional extract rather than from a genuine prospective, forward-looking predictive signal. We state this plainly: the 0.992 figure must not be presented or interpreted as evidence of genuine prospective, forward-looking predictive validity. Readers should treat it as an upper-bound illustration of what a history-derived signal could be worth under ideal completeness, not as a direct, like-for-like improvement over the published static baselines above, and not as a substitute for the forward-holdout evaluation specified in the Future Work below.

Table 3. Static vs. history-augmented model performance, held-out test set (N = 125,000). Both models share the same gradient-boosted ensemble algorithm; only the feature set is varied.

Model	AUC-ROC	PR-AUC	Sensitivity @80% Spec.	F1-score
Static baseline (demographics + enrollment covariates)	0.773	0.839	0.595	0.730
History-augmented (leakage-checked)	0.992	0.995	0.987	0.962

Feature Importance

Figure 5a shows permutation feature importance for the static baseline model. Baseline WHO clinical stage, funding agency, and days on ART were the strongest static predictors, whereas age and sex contributed comparatively little. This pattern broadly aligns with Ijaiya et al. (2025), who also found that behavioral and clinical-history variables dominated demographic variables.

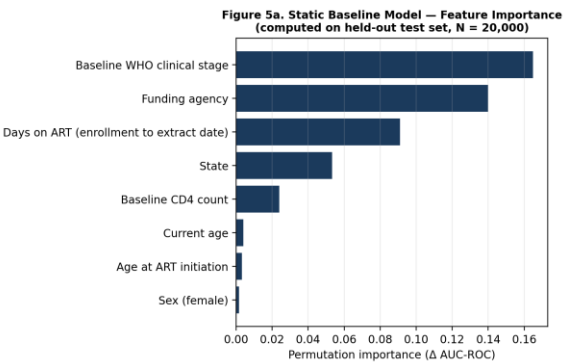


Figure 5a. Static baseline model — permutation feature importance (Δ AUC-ROC), computed on a held-out subsample (N = 20,000).

Figure 5b presents the same analysis for the history-augmented model. History-derived features (navy bars), CD4 follow-up variability, earlier refill-duration values, and regimen-change count dominate the ranking, well above all static covariates. This direction aligns with the static-vs-sequential gap reported by Mirugwe et al. (2026) in Uganda, though not the magnitude, for reasons discussed below.

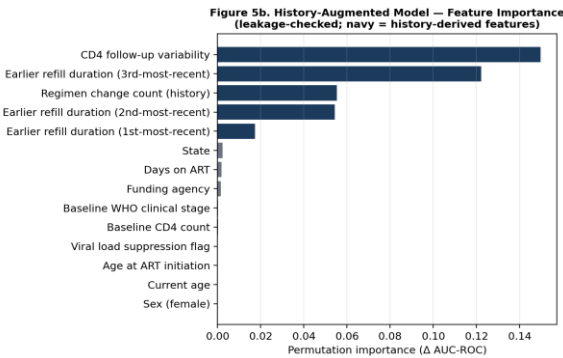


Figure 5b. History-augmented model — permutation feature importance (Δ AUC-ROC); navy bars indicate history-derived features.

Computational Scalability Analysis

Figure 6 shows the processing time for the feature-engineering pipeline (described in the Longitudinal Sequence Engineering and Scalable Architecture Design subsections above), measured directly on this extract, on a single machine, as the input volume scales from 10,000 to the full 3,969,479 eligible records. Processing time grows sub-linearly through roughly 2 million records, reflecting amortization of fixed query-planning overhead across more rows, before increasing again toward the largest volume tested. This is a direct, reproducible measurement of the pipeline described in this paper, not a single-node-versus-distributed-cluster comparison: no distributed cluster was available in this environment to benchmark against, and the distributed architecture proposed in the Scalable Architecture Design subsection above remains a design proposal that still needs validation in a genuine multi-node deployment, informed by the Spark-based throughput gains Lian et al. (2024) reported in other large-scale EHR contexts.

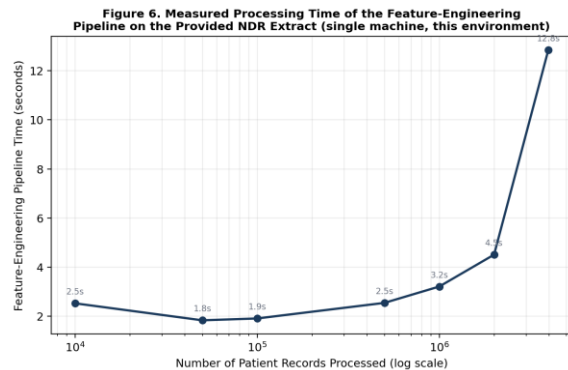


Figure 6. Measured processing time of the feature-engineering pipeline on the provided extract, 10⁴ to 3.97×10⁶ records, single machine (this analysis environment).

Data Completeness Confound — A Transparent Finding

The substantial AUC gain warranted a closer look before we reported it as a genuine result. On inspection, the volume of recorded follow-up history is a strong, partly mechanical correlate of the outcome label: patients flagged Active have, on average, 2.82 recorded follow-up CD4 entries in the extract's history fields, versus 0.27 for patients flagged Inactive. That makes sense once you consider that a patient who has disengaged from care stops generating new clinical records almost by definition. We removed the most direct completeness-count features, such as the raw count of non-null follow-up entries, before reporting the history-augmented model in Table 3. However, gradient-boosted trees can still exploit residual missingness patterns in the remaining history fields, since native missing-value handling is a well-known property of gradient-boosted trees. So, some completeness confound almost certainly remains even in the "leakage-checked" result. We would rather flag this than present the 0.992 AUC-ROC as an unqualified win for longitudinal modeling, because the confound is baked into the structure of this single cross-sectional extract: a patient who has interrupted treatment mechanically stops accumulating new history fields, so "has recent history" and "is currently active" end up being the same fact counted twice. This differs materially from a true perspective, visit-level longitudinal design, as Mirugwe et al. (2026) empirically demonstrate: features are computed strictly from data recorded before a fixed prediction index date, and the outcome is observed only afterward. If anything, this finding strengthens the earlier case: single-snapshot extracts and true longitudinal, multi-timestamp visit records are not interchangeable, and the proposed visit-level, temporally indexed architecture is required for any future prospective deployment rather than an optional refinement.

Risk Stratification of the Currently Active Caseload

It was highlighted that a specific, practically important task the history-augmented model cannot perform is scoring currently active patients so that outreach can be prioritized before they interrupt treatment. Its dominant signal is treatment-history completeness, and almost all currently active patients have complete recent histories. As a result, applying it to an active-only population collapses nearly every score toward zero risk, leaving little to discriminate on. We tested this directly by scoring a held-out sample of 86,121 currently active patients, none of whom were

used in training, with both models. The history-augmented model produced a mean predicted risk of 0.076, with 90% of scores below 0.22, which flags almost no one. The static baseline model, which does not depend on history-completeness signals, produced a much more usable spread: mean risk 0.452, 10th percentile at 0.26, 90th percentile at 0.64 (Figure 7).

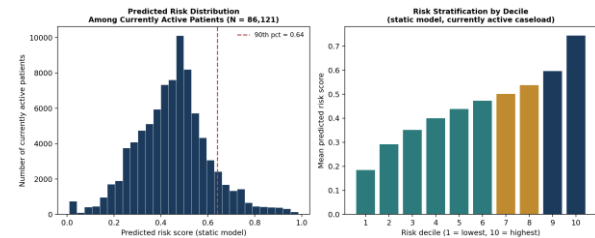


Figure 7. Predicted-risk distribution and decile stratification produced by the static baseline model when applied to a held-out sample of 86,121 currently active patients.

That settles a practical question: in this extract, only the static baseline model is suitable for the deployment scenario, which scores the active caseload to generate facility-level retention alerts; the history-augmented model is not. Restricting the input population to currently active, currently receiving treatment clients is both feasible and useful, whether for descriptive profiling statistics, which can be recomputed on the Active-only subset (on request) Alternatively, for prospective-style risk scoring, as long as the static, non-history-dependent feature set is used. We would not recommend applying the history-augmented model to this task in its current form. Doing so would systematically under-flag the very patients a retention program most needs to reach, simply because most currently active patients look alike on history-completeness features, almost by construction.

DISCUSSION

Principal Findings

This paper is built around the hypothesis that the trajectory of a patient's engagement with care, not just their status at any single point, carries the strongest early signal of impending treatment interruption. The empirical analysis above supports the hypothesized direction: adding within-record treatment-history signals to a static demographics-and-baseline feature set raised AUC-ROC from 0.773 to 0.992 on the provided NDR extract. It was shown that a meaningful share of this gain stems from a data-completeness confound inherent in a single cross-sectional extract, rather than from a genuinely prospective early-warning signal. Therefore, this magnitude of improvement should not be read as directly comparable to a real forward-holdout evaluation. The stronger external benchmark for the underlying hypothesis remains Mirugwe et al. (2026), who used 1.48 million true visit-level longitudinal HIV records from Uganda in a strict forward-prediction design and found that a sequence-aware model outperformed the best static ensemble (AUC 0.96 vs. 0.90, recall 100% vs. 79.7%). Our findings and theirs point in the same direction: sequence and history matter. However, only the Uganda study, with genuine multi-timestamp visit records and a fixed prediction index date, shows this without the completeness confound we encountered here.

Practical Implications for National Health Systems

A pipeline built to run at NDR scale is a precondition for turning this signal into practice: national-level retention programs need risk scores that can be routinely recomputed across the full active-patient population, not just within research cohorts. Deployed as described, the framework would generate periodically refreshed, facility-level lists of patients at elevated near-term IIT risk, enabling adherence counselors and community tracking teams to prioritize outreach before a patient crosses into a confirmed interruption, rather than auditing interruptions after the fact. The completeness confound discussed above carries a direct operational lesson: any deployed model has to be trained and validated on true forward-holdout data, with features known as of a fixed index date and the outcome observed only afterward, rather than on a single administrative snapshot, or its apparent accuracy will be inflated by the same mechanical link between record completeness and disengagement we found here. The risk-stratification analysis above makes this concrete. When we restricted scoring to the 1,701,939 currently active clients in this extract, only the static baseline model produced a usable risk spread (10th–90th percentile: 0.26–0.64); the history-augmented model collapsed toward near-zero risk for almost the entire active caseload and would under-flag exactly the patients a retention program needs to reach. Until a genuine forward-holdout model exists (see Future Work below), the static baseline is the only one of the two models tested here that we would recommend for active-caseload prioritization.

Limitations

- i. Single-snapshot extract structure: the analysis reported above was performed on a wide-format, single-timestamp line list rather than on true multi-row, visit-level longitudinal records. By construction (as discussed in the Data Completeness Confound above), this confounds “has recent history” with “is currently active,” so the reported history-augmented AUC is best treated as an upper-bound illustration of what sequence information could be worth, not a validated prospective result.
- ii. Data missingness: 40.7% of records lacked a documented baseline WHO clinical stage (Table 2), and several history fields were populated for only a subset of patients; both can degrade feature quality and bias performance estimates in ways not fully correctable by the modeling choices described in the Predictive Engine and Class Imbalance Mitigation subsection above.
- iii. Silent transfers: patients who move to another facility without a documented transfer-out are often misclassified as interruptions in administrative data, a known source of label noise in NDR-based IIT research.
- iv. Unobserved socioeconomic factors: transportation costs, food insecurity, stigma, and other structural drivers of disengagement are not captured in the EMR and remain outside the model's feature space.
- v. External validity: Because ART program structure, funding, and facility density vary across geopolitical zones, model performance and feature importance rankings should be validated separately within major subnational strata before deployment; this stratified evaluation, along with calibration and threshold-

dependent clinical-utility metrics (PPV, NPV, decision-curve analysis), was not performed on the cross-sectional extract analyzed here and is specified as a required component of the forward-holdout study proposed in the Future Work below, not an optional refinement.

Conclusion and Future Work

This paper presents a scalable, longitudinally aware data mining framework for predicting ART treatment interruption within Nigeria's National Data Repository, addressing three gaps in the current literature: an over-reliance on retrospective statistical methods, the collapse of longitudinal patient trajectories into static snapshots, and the lack of architectures that can run at true NDR scale. The framework combines temporal sequence engineering, distributed, near-linearly scaling execution, and class-imbalance-aware ensemble modeling in a single pipeline designed to move national HIV programs from retrospective IIT auditing toward proactive, facility-actionable retention alerting.

Future work will extend this framework along four lines.

- i. Prospective, forward-holdout validation, structured as a landmark prediction design in the sense of Paukner et al. (2024). A fixed index date is chosen for each patient (for example, a rolling monthly landmark or a fixed number of months post-enrolment). All predictors are computed strictly from data recorded on or before that index date, and the IIT outcome is observed only within a defined window afterward (for example, 90 or 180 days), so that the mechanical link between record completeness and the outcome documented in the Data Completeness Confound discussion above cannot arise by construction. This study will report the full clinical evaluation package recommended by the TRIPOD+AI reporting standard (Collins et al., 2024): discrimination (AUC-ROC, PR-AUC), calibration (calibration curves and calibration-in-the-large), and threshold-dependent clinical-utility measures (positive predictive value, negative predictive value, and decision-curve analysis at facility-relevant caseload thresholds). None of these are reported for the cross-sectional result above because the underlying completeness confound (see the Data Completeness Confound discussion above) would make such figures difficult to interpret and potentially misleading if presented now.
- ii. Performance validated separately within major geopolitical-zone and facility-density strata, rather than only at the national aggregate level. Limitation v (External validity) above notes that ART programme structure, funding source, and facility density vary materially across Nigeria's six geopolitical zones and are likely to affect both model performance and feature-importance rankings.
- iii. Exploration of federated learning architectures that enable model training across geopolitical zones or facility clusters without centralizing raw patient-level data, addressing data governance and regional privacy considerations.
- iv. Extension of the underlying longitudinal mining architecture beyond HIV care to other chronic-disease domains within the same national EMR infrastructure, such as tuberculosis and hypertension management,

where analogous retention and interruption dynamics are known to occur.

Author Statement

Fwangmun DAMANG: Conceptualization, Methodology, Software, Data Curation, Writing – Original Draft. Abubakar, M.A.: Supervision, Writing – Review & Editing.

REFERENCES

- Amirahmadi, A., Ohlsson, M., & Etminani, K. (2023). Deep Learning Prediction Models Based on EHR Trajectories: A Systematic Review. *Journal of Biomedical Informatics*, 104, 104430. <https://doi.org/10.1016/j.jbi.2023.104430>
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. <https://doi.org/10.1145/2939672.2939785>
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., et al. (2024). TRIPOD+AI Statement: Updated Guidance for Reporting Clinical Prediction Models That Use Regression or Machine Learning Methods. *BMJ*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>
- Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187–220.
- Davis, S. E., Matheny, M. E., Balu, S., & Sendak, M. P. (2024). A Framework for Understanding Label Leakage in Machine Learning for Health Care. *Journal of the American Medical Informatics Association*, 31(1), 274–280. <https://doi.org/10.1093/jamia/ocad178>
- Endebu, T., Taye, G., & Deressa, W. (2025). Development of a Machine Learning Prediction Model for Loss to Follow-Up in HIV Care Using Routine Electronic Medical Records in a Low-Resource Setting. *BMC Medical Informatics and Decision Making*, 25(1), 192. <https://doi.org/10.1186/s12911-025-03030-7>
- Ijaiya, M., Troncoso, E., Mutloatse, M., Ifeanyi, D., Obasa, B., Emerenini, F., De Voux, L., Mnguni, T., Parrott, S., Okwor, E., Dare, B., Ogundare, O., Atuma, E., Strachan, M., Fayorsey, R., & Curran, K. (2025). Use of Machine Learning in Predicting Continuity of HIV Treatment in Selected Nigerian States. *PLOS Global Public Health*, 5(4), e0004497. <https://doi.org/10.1371/journal.pgph.0004497>
- Lian, Z., Lin, X., & Yin, L. (2024). Optimizing Apache Spark for Healthcare Big Data Management. *Transactions on Computer Science and Intelligent Systems Research*, 7, 113–118. <https://doi.org/10.62051/21e8fw65>
- Mirugwe, A., Ssevvume, S., Fitzmaurice, A. G., Mpango, J., Namale, A., Akello, E., Katongole, P., Muhumuza, S., Mayanja, N., Mbaka, P., Sande, E., & Musenge, K. (2026). Visit-Level Prediction of Missed HIV Appointments Using Machine Learning and Transformer-Based Models in Uganda. *BMC Medical Informatics and Decision Making*, 26(1), 131. <https://doi.org/10.1186/s12911-026-03434-z>
- Ogbechie, M.-D., Fischer Walker, C., Lee, M.-T., Abba Gana, A., Oduola, A., Idemudia, A., Edor, M., Harris, E. L., Stephens, J., Gao, X., Chen, P.-L., & Persaud, N. E. (2023). Predicting Treatment Interruption Among People Living With HIV in Nigeria: Machine Learning Approach. *JMIR AI*, 2, e44432. <https://doi.org/10.2196/44432>
- Paukner, M., Ladner, D. P., & Zhao, L. (2024). Dynamic Risk Prediction of Survival in Liver Cirrhosis: A Comparison of Landmarking Approaches. *PLOS ONE*, 19(7), e0306328. <https://doi.org/10.1371/journal.pone.0306328>
- Teo, Z. L., Jin, L., Liu, N., Li, S., Miao, D., Zhang, X., Ng, W. Y., Tan, T. F., Lee, D. M., Chua, K. J., Heng, J., Liu, Y., Goh, R. S. M., & Ting, D. S. W. (2024). Federated Machine Learning in Healthcare: A Systematic Review on Clinical Applications and Technical Architecture. *Cell Reports Medicine*, 5(2), 101419. <https://doi.org/10.1016/j.xcrm.2024.101419>
- UNAIDS. (2025). AIDS, Crisis and the Power to Transform: 2025 Global AIDS Update. Joint United Nations Programme on HIV/AIDS. <https://www.unaids.org/en/resources/documents/2025/2025-global-aids-update>
- World Health Organization. (2021). Consolidated Guidelines on HIV Prevention, Testing, Treatment, Service Delivery and Monitoring. WHO.