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.

# Feature leakage inflates machine-learning estimates of antiretroviral non-adherence risk in children living with HIV: a tiered sensitivity analysis of the EARTH cohort

**Ssegawa Abdulkarim**
Uganda Technology and Management University (UTAMU)
**Ahimbisibwe Mbabazi Mordecai**
Uganda Technology and Management University (UTAMU)
**Joseph Igaga**
Uganda Technology and Management University (UTAMU)
**Denis Bbosa**
Mbarara University of Science and Technology (MUST)
**Fred Paul Mark Jjunju**
fpmjjunju@utamu.ac.ug

Uganda Technology and Management University (UTAMU)

**Short Report**

**Keywords:** children living with HIV, antiretroviral therapy, adherence, machine learning, data leakage, sensitivity analysis, sub-Saharan Africa, EARTH cohort

**Posted Date:** November 11th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-11163970/v1

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

**Additional Declarations:** No competing interests reported.

## Abstract

# Objective

A companion analysis of the EARTH cohort developed machine-learning models to identify suboptimal antiretroviral therapy adherence in children living with HIV after excluding adherence-process fields that directly contributed to the outcome definition. This Research Note quantifies how retaining those fields changes apparent model performance. Using a separate sensitivity-analysis extract from the same cohort, we compared three nested predictor tiers across 590 visits from 208 children: leakage-naive, loosely controlled, and strictly controlled, using fourteen classifiers and grouped five-fold cross-validation.

# Results

Including outcome-defining adherence-process fields markedly inflated discrimination across all fourteen classifiers, increasing AUROC by 0.24–0.41 and AUPRC by 0.33–0.68; six classifiers reached AUROC 1.000. After removing these fields, CatBoost achieved AUROC 0.735. Excluding an additional ART treatment-interruption variable, which dominated permutation importance despite not formally defining the outcome, reduced CatBoost AUROC to 0.680 and AUPRC to 0.407, closely matching the companion analysis. The findings show that adherence-process variables can create both direct and near-label leakage, producing deceptively strong performance. Tiered leakage auditing should therefore accompany machine-learning studies that predict adherence from routinely collected process data, particularly when predictors overlap conceptually with outcome construction before interpreting apparent performance as clinically useful predictive signal.

## Introduction

Machine-learning models for antiretroviral therapy (ART) adherence and retention are increasingly being developed in low-resource settings [1–3]. A recurrent methodological hazard is data leakage: predictors may contain information used directly or indirectly to construct the target label, causing apparent performance to reflect outcome construction rather than genuine predictive signal [4, 5].

A companion analysis of the EARTH cohort excluded adherence-process fields that directly contributed to its four-criterion adherence outcome before model training. That analysis reported more modest discrimination but did not quantify how much apparent performance would increase if those fields were retained. This Research Note reports that sensitivity analysis using a separately prepared extract from the same cohort and study window. It is a side product of the primary modelling study rather than a new data-collection exercise.

## Main text

# Data, cohort reconstruction and outcome

The sensitivity extract contained 1,240 REDCap rows and 107 columns, including genuine follow-up visits and non-visit administrative events. Because the extract did not contain an explicit participant identifier, participant blocks were reconstructed from the strictly increasing order of REDCap event codes (0, 1, 2, 3, 6, 7, 8), beginning a new participant whenever the code failed to increase relative to the preceding row. This produced 208 children with at least one genuine follow-up visit, matching the companion analysis population count. Restricting the data to genuine follow-up visits (Visit Number 1, 2 or 3; excluding administrative rows coded 5) yielded 590 visit-level records.

Suboptimal adherence was reconstructed from the available raw fields using the companion study definition: caregiver-reported missed dose since the previous visit, pharmacy collection delay exceeding 10% of the inter-visit interval, reported running out of medicine, or an implausibly large amount of medicine remaining. The logical-OR rule reproduced the extract's precomputed outcome for all 590 visits and yielded 157 suboptimal-adherence visits (26.6%).

## Three nested predictor tiers

Thirty administrative or structural fields were removed from all analyses, leaving 76 candidate predictors. Tier 1 (leakage-naive; 76 features) retained outcome-defining adherence-process fields, including missed-dose report, cumulative pharmacy-delay days, inter-visit interval used to calculate the delay fraction, running out of medicine, and medicine remaining. Tier 2 (loosely controlled; 71 features) removed these five fields, mirroring the companion analysis feature-control strategy. Tier 3 (strictly controlled; 70 features) additionally removed ART treatment interruptions since the previous visit, a field that did not formally define the outcome but behaved as a near-leakage proxy. The extract lacked a study-site field available in the companion manuscript, so Tier 2 was not a byte-for-byte reproduction of that feature set.

## Model validation

The same modelling pipeline was applied to each tier. Fourteen classifiers were evaluated: logistic regression, linear and quadratic discriminant analysis, Gaussian naive Bayes, k-nearest neighbours, radial-basis-function support vector machine, decision tree, random forest, gradient boosting, AdaBoost, XGBoost, LightGBM, CatBoost, and a multilayer perceptron. Validation used stratified grouped five-fold cross-validation with children as the grouping unit and in-fold SMOTE oversampling. A classification threshold was selected to achieve sensitivity of at least 0.85.

AUROC and AUPRC were calculated from pooled out-of-fold predictions. Grouped bootstrap resampling over children (1,000 resamples) was used for 95% confidence intervals for selected models. CatBoost

permutation importance was estimated as the mean decline in held-out AUPRC across 20 shuffles on a grouped 70/30 split.

## Results

In Tier 1, six classifiers reached AUROC 1.000 and AUPRC 1.000, while several others were also near-perfect (Table 1; Fig. 1). For CatBoost, the largest reported permutation-importance values were medicine remaining (0.454), missed doses (0.342), pharmacy-delay days (0.036), and running out of medicine (0.030), indicating that the model was largely reading the outcome through its defining components (Fig. 2).

Removing the direct outcome-defining fields in Tier 2 reduced performance markedly. AdaBoost achieved AUROC 0.752 and AUPRC 0.600; CatBoost achieved AUROC 0.735 and AUPRC 0.566. The remaining ART treatment-interruption field then dominated CatBoost permutation importance (0.134), compared with 0.006 for the next reported predictor, breastfeeding status, suggesting a further near-leakage channel.

After removing ART treatment interruption in Tier 3, CatBoost fell to AUROC 0.680 and AUPRC 0.407. Across all fourteen classifiers, retaining direct adherence-process fields increased AUROC by 0.237–0.407 and AUPRC by 0.326–0.681. Full model-level results are provided in Additional file 1.

Table 1
Benchmark of seven classifiers across the three predictor tiers

| Model | Tier 1 AUROC | Tier 1 AUPRC | Tier 2 AUROC | Tier 2 AUPRC | Tier 3 AUROC | Tier 3 AUPRC |
|---|---|---|---|---|---|---|
| AdaBoost | 1.000 | 1.000 | 0.752 | 0.600 | 0.705 | 0.442 |
| CatBoost | 1.000 | 1.000 | 0.735 | 0.566 | 0.680 | 0.407 |
| Gradient boosting | 1.000 | 1.000 | 0.733 | 0.573 | 0.672 | 0.418 |
| Random forest | 1.000 | 0.999 | 0.727 | 0.516 | 0.663 | 0.370 |
| XGBoost | 1.000 | 1.000 | 0.711 | 0.511 | 0.625 | 0.368 |
| LightGBM | 1.000 | 1.000 | 0.706 | 0.535 | 0.631 | 0.366 |
| Logistic regression | 0.986 | 0.984 | 0.674 | 0.485 | 0.601 | 0.351 |

*Tier 1: leakage-naive (76 features). Tier 2: direct outcome-defining adherence-process fields removed (71 features). Tier 3: ART treatment-interruption field additionally removed (70 features). Values are pooled out-of-fold estimates from grouped five-fold cross-validation. Complete Tier 1 and Tier 2 fourteen-classifier results and the seven reported Tier 3 models are in Additional file 1.*

## Interpretation

The sensitivity analysis demonstrates how label construction can dominate apparent machine-learning performance. Tree-based models were particularly affected because the adherence outcome is a logical OR of thresholded process variables, a structure that tree ensembles can represent efficiently. The contrast between near-perfect tree-ensemble performance and weaker linear or distance-based models is therefore a useful diagnostic signal, although it cannot replace direct feature auditing.

Adherence-process variables are not intrinsically invalid predictors. Pharmacy refill or missed-dose information may be appropriate when predicting a conceptually distinct outcome such as viral suppression. The concern arises when the target itself is constructed from those same measurements or close proxies. For such studies, a tiered analysis that sequentially removes direct outcome components and plausible near-proxies can reveal how much apparent discrimination depends on leakage rather than transportable signal.

# Limitations

- • The analysis used one cohort and one outcome definition, so the magnitude of inflation should not be interpreted as a universal correction factor for adherence-prediction studies.
- • The sensitivity extract lacked a study-site field present in the companion analysis and contained 590 rather than 592 follow-up visits; Tier 2 therefore cannot be treated as an exact replication of the companion feature matrix.
- • Participant identifiers were reconstructed from REDCap event ordering rather than supplied directly. The resulting participant count matched the companion analysis, but rare misassignment cannot be excluded.
- • The classification of ART treatment interruption as near-leakage is an exploratory interpretation based on its dominant permutation importance and conceptual proximity to adherence; it is not a formal proof that every remaining predictor is leakage-free.
- • No hyperparameter tuning, external validation, or prospective validation was performed. The results therefore quantify sensitivity to predictor-set construction within this cohort rather than clinical readiness.

## Abbreviations

ART
antiretroviral therapy
AUPRC
area under the precision-recall curve
AUROC
area under the receiver operating characteristic curve
CI

confidence interval
CLHIV
children living with HIV
EARTH
Early Antiretroviral Therapy and child Health cohort
ML
machine learning
SMOTE
synthetic minority over-sampling technique.

## Declarations

# Ethics approval and consent to participate

This is a secondary analysis of data collected under the EARTH cohort study. Ethical approvals waiver and the permission use the data from EARTH project was obtained prior to the study (see additional file for more details) and the REC committed adhered to the Declaration of Helsinki. The parent study obtained written informed consent from each child's parent or legal guardian for participation and secondary use of de-identified data. Before submission, the authors should insert the exact names of the approving ethics committees/institutional review boards and the relevant approval or reference numbers from the parent EARTH study records.

# Consent for publication

Not applicable. This manuscript contains no identifiable individual-level data, images, or videos.

# Competing interests

The authors declare that they have no competing interests.

# Additional files

Supplementary benchmark tables and figures for the EARTH cohort leakage-sensitivity analysis. It contains the fourteen-classifier Tier 1 and Tier 2 benchmark tables, the seven reported Tier 3 models, a data-preparation and predictor-tier flow diagram, and a CatBoost performance summary across tiers.

# Funding

No funding

# Author Contribution

All authors contributed equally and read and approved the final manuscript.

# Acknowledgement

The authors thank UTAMU, MUST and the EARTH cohort investigators, clinical teams, participating children, and their families for the conducive environment.

# Data Availability

The datasets analysed in this study are not publicly available because of participant confidentiality and institutional data-sharing agreements governing the EARTH cohort. Data may be available from the corresponding author and the EARTH cohort study team on reasonable request and subject to applicable approvals. Analysis code is available from the corresponding author on reasonable request. Summary tables and figures supporting this Research Note are included in SI.

## References

1. Najjuuko C, Brathwaite R, Xu Z, Kizito S, Lu C, Ssewamala FM. Using machine learning to predict poor adherence to antiretroviral therapy among adolescents living with HIV in low resource settings. AIDS. 2025;39(9):1204–13. 10.1097/QAD.0000000000004163.
2. Mirugwe A, Ssevvume S, Fitzmaurice AG, et al. Visit-level prediction of missed HIV appointments using machine learning and transformer-based models in Uganda. BMC Med Inf Decis Mak. 2026;26:131. 10.1186/s12911-026-03434-z.
3. Lugoba MD, Sangeda RZ, Mutagonda RF, Mwakyomo J, Musiba G, Sambu V et al. Leveraging machine learning models and pharmacy refill adherence as a cost-effective proxy for predicting HIV viral suppression during antiretroviral therapy in resource-limited settings. medRxiv. 2026;2026.01.05.26343496.
4. Kaufman S, Rosset S, Perlich C, Stitelman O. Leakage in data mining: formulation, detection, and avoidance. ACM Trans Knowl Discov Data. 2012;6(4):15.
5. Hossain M, Kibria N, Shahriar F. Evaluating reliability in machine learning models for early chronic kidney disease prediction: a systematic review of data leakage and predictor stability. arXiv. 2026;arXiv:260711963.
6. Tagarro A, Domínguez-Rodríguez S, et al. High mortality following early initiation of antiretroviral therapy in infants living with HIV from three African countries. eClinicalMedicine. 2024;73:102648.

10.1016/j.eclinm.2024.102648.

# Figures

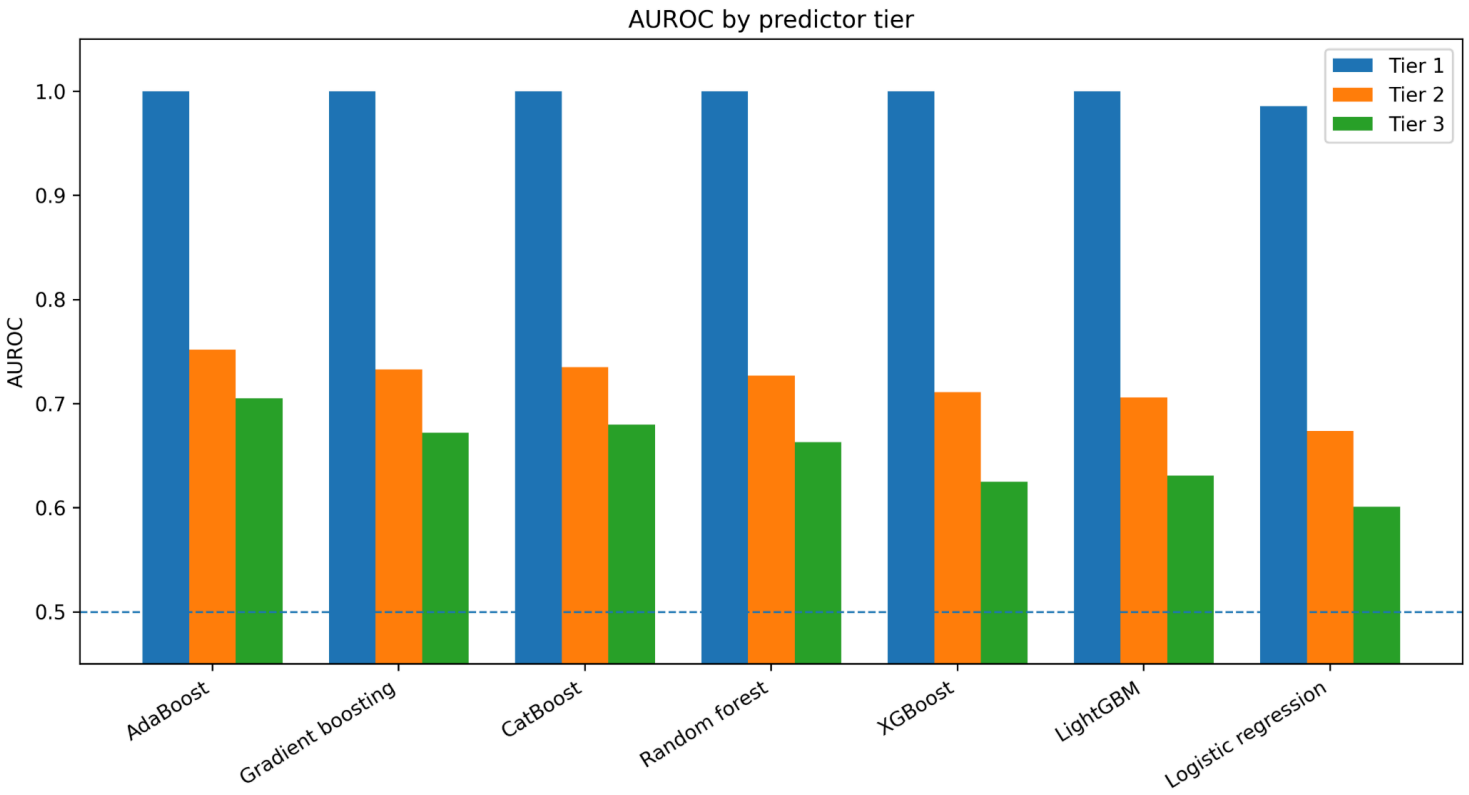



**Figure 1**

**AUROC by feature-set tier for seven classifiers.** Bars show pooled out-of-fold AUROC under leakage-naive, loosely controlled, and strictly controlled predictor sets. The dashed line marks chance-level discrimination (AUROC = 0.5).

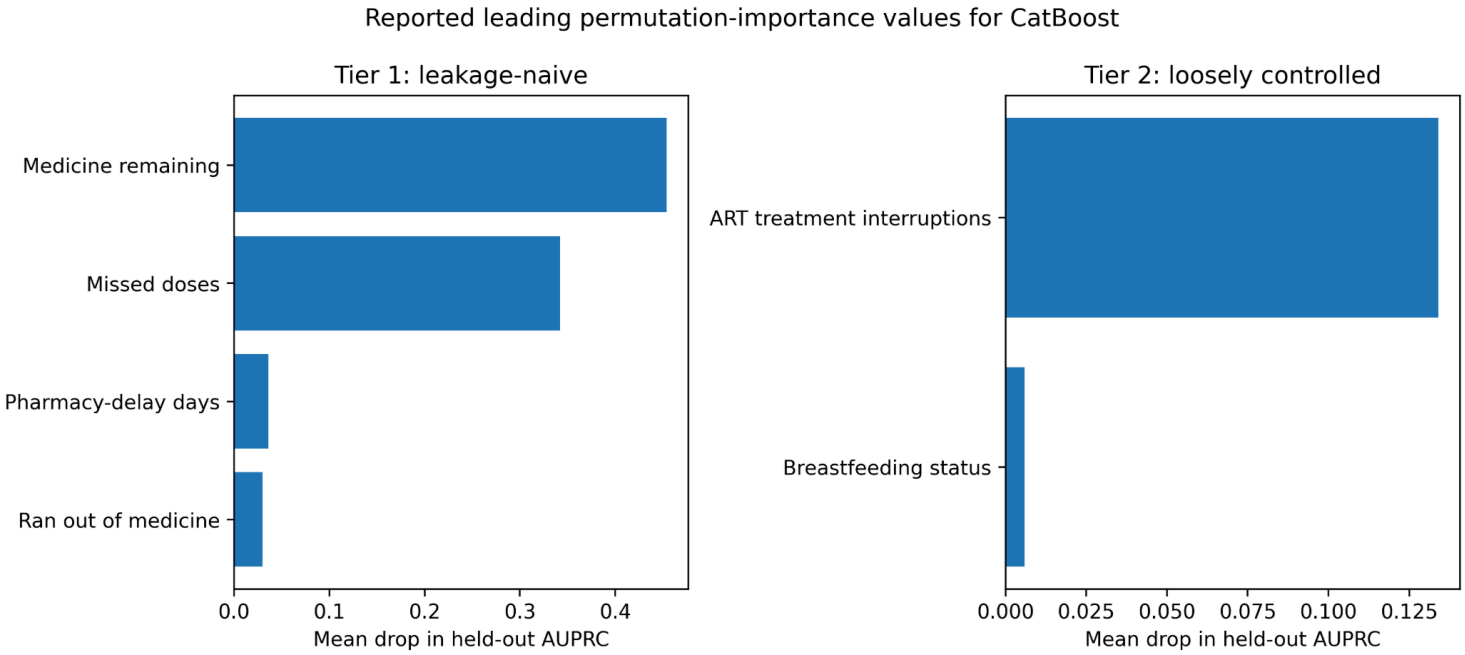



**Figure 2**

**Reported leading permutation-importance values for CatBoost.** Mean decline in held-out AUPRC after feature shuffling (20 repeats) on a grouped 70/30 split. Only predictor values explicitly reported in the source analysis are shown.

# Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- EARTHLeakageSI.docx