Title

Prevention of adverse HIV treatment outcomes: machine learning to enable proactive support of people at risk of HIV care disengagement in Tanzania.

Journal

BMJ Open, 14(9)

Authors

Xie, Zhongming
Hu, Huiyu
Kadota, Jillian
et al.

Publication Date

2024-09-24

DOI

10.1136/bmjopen-2024-088782

Peer reviewed

BMJ Open

Prevention of adverse HIV treatment outcomes: machine learning to enable proactive support of people at risk of HIV care disengagement in Tanzania

Zhongming Xie,[1] Huiyu Hu ,[1] Jillian L Kadota,[1] Laura J Packel,[1] Matilda Mlowe,[2] Sylvester Kwilasa,[3] Werner Maokola,[3] Siraji Shabani,[3] Amon Sabasaba,[2] Prosper F Njau,[3] Jingshen Wang,[1] Sandra I McCoy [1]

To cite: Xie Z, Hu H, Kadota JL, *et al*. Prevention of adverse HIV treatment outcomes: machine learning to enable proactive support of people at risk of HIV care disengagement in Tanzania. *BMJ Open* 2024;**14**:e088782. doi:10.1136/bmjopen-2024-088782

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2024-088782).

ZX and HH are joint first authors. JW and SIM are joint senior authors.

Received 16 May 2024
Accepted 27 August 2024

[1]School of Public Health, University of California, Berkeley, California, USA
[2]Health for a Prosperous Nation, Dar es Salaam, Tanzania, United Republic of
[3]United Republic of Tanzania Ministry of Health, Dodoma, Tanzania, United Republic of

Correspondence to
Huiyu Hu;
huiyuhu@berkeley.edu

ABSTRACT

Objectives This study aimed to develop a machine learning (ML) model to predict disengagement from HIV care, high viral load or death among people living with HIV (PLHIV) with the goal of enabling proactive support interventions in Tanzania. The algorithm addressed common challenges when applying ML to electronic medical record (EMR) data: (1) imbalanced outcome distribution; (2) heterogeneity across multisite EMR data and (3) evolving virological suppression thresholds.

Design Observational study using a national EMR database.

Setting Conducted in two regions in Tanzania, using data from the National HIV Care database.

Participants The study included over 6 million HIV care visit records from 295 961 PLHIV in two regions in Tanzania's National HIV Care database from January 2015 to May 2023.

Results Our ML model effectively identified PLHIV at increased risk of adverse outcomes. Key predictors included past disengagement from care, antiretroviral therapy (ART) status (which tracks a patient's engagement with ART across visits), age and time on ART. The downsampling approach we implemented effectively managed imbalanced data to reduce prediction bias. Site-specific algorithms performed better compared with a universal approach, highlighting the importance of tailoring ML models to local contexts. A sensitivity analysis confirmed the model's robustness to changes in viral load suppression thresholds.

Conclusions ML models leveraging large-scale databases of patient data offer significant potential to identify PLHIV for interventions to enhance engagement in HIV care in resource-limited settings. Tailoring algorithms to local contexts and flexibility towards evolving clinical guidelines are essential for maximising their impact.

STRENGTHS AND LIMITATIONS OF THIS STUDY

⇒ Addressing large-scale electronic medical record (EMR) data challenges: We used a national EMR database to effectively identify at-risk HIV patients, managing issues like imbalanced outcome distribution, data heterogeneity and evolving clinic guidelines, showing the robustness and applicability of the machine learning model to other health areas.
⇒ Participatory data science approach: Collaboration among clinicians, academics and policy-makers ensured the relevance and applicability of the findings for real-world use.
⇒ Data quality and timeliness: The limitations include reliance on periodic data downloads (not real-time), all-cause mortality data (not HIV-specific) and inherent EMR data issues (incompleteness and inaccuracies), which may affect model performance and applicability.

INTRODUCTION

The Joint United Nations Programme on HIV/AIDS (UNAIDS) has set forth ambitious '95-95-95' targets for 2030, aiming for 95% of people living with HIV (PLHIV) to know their status, 95% of those diagnosed to receive antiretroviral therapy (ART) and 95% of those on ART to achieve suppressed viral loads.[1] Central to these goals is lifelong effective ART use, which improves the length and quality of life among those living with HIV and reduces the risk of transmission.[2 3] Despite significant progress in scaling up access to ART, disengagement from care and poor ART adherence remain persistent challenges in sub-Saharan Africa.[4 5] Therefore, methods to further enhance engagement in HIV care would optimise the clinical benefits of ART and its public health preventive impact.

Efforts to improve engagement, such as enhanced adherence counselling in response to high viraemia,[6] visit text reminders,[7] financial incentives for clinic attendance,[8] location tracking,[9] multimonth dispensing[10] and differentiated care,[11] have been implemented with mixed results regarding effectiveness, scalability and efficiency. For example, a recent randomised trial demonstrated

that small, short-term financial incentives improved viral suppression 6 months after ART initiation by 13 percentage points.[8] Despite this benefit, 73% of the comparison group achieved viral suppression without the intervention, indicating the potential for greater impact through more targeted resource allocation. In the last decade, the increasing digitisation of medical records in sub-Saharan Africa has provided opportunities to use machine learning (ML) to enhance HIV care delivery,[12 13] including predicting virological outcomes and future missed visits.[14–16] ML methods are particularly useful in this big data context due to their ability to handle complex, non-linear relationships and interactions within the data as well as manage a large number of highly interactive combinations of predictors, which traditional regression models might miss.[17] Building on this premise, a previous study revealed the capacity of ML methods to harness both digitised medical records and research study data, predicting the likelihood of patients disengaging from HIV care in limited-resource settings,[18] highlighting the need for further exploration of this innovative approach.

However, the secondary use of electronic medical record (EMR) data is not without challenges. A primary concern is the potentially limited reliability of prediction models in real-world environments, due to the multifaceted, varied nature of patient data.[19] Issues such as missing data, irregular visit dates and the heterogeneity of patient records can complicate model training and accuracy,[20 21] which could theoretically lead to inaccuracies in patient care decisions, inappropriate treatments or resource misallocation. Furthermore, the literature seldom addresses the trade-offs between the development of multiple algorithms tailored to specific regional characteristics vs a 'universal' model for all contexts. Another emerging challenge is the dynamic nature of clinical and laboratory policies and recommendations. For example, within the context of HIV, the WHO recommends virological monitoring using a viral suppression threshold of 1000 copies/mL. However, increasingly, lower thresholds are being used at the programmatic level, such as 50 copies/mL.[22] This requires data scientists to make decisions on how to select and define explanatory factors and outcomes in their algorithms against a backdrop of changing clinical and monitoring guidelines.

We aim to use ML with data from Tanzania's National HIV Care and Treatment programme to identify PLHIV at increased risk of adverse outcomes, addressing key challenges such as managing imbalanced data, choosing between universal and site-specific algorithms, and adapting to changes in viral load thresholds for suppression. Our goal is to identify individuals accessing care at two health facilities in Tanzania who are at higher risk of HIV care disengagement, high viral load or death and enrol these individuals into a future intervention study aimed at improving retention and virological suppression outcomes, thereby offering a direct, tangible pathway towards proactive intervention to support lifelong engagement in HIV care.

METHODS

Objectives

The primary objective of the study was to construct and validate an ML algorithm capable of generating individualised risk scores for potential future care disengagement or other adverse outcomes among PLHIV using routine clinical and pharmaceutical data from Tanzania's HIV Care and Treatment database. The algorithm aims to address common EMR data challenges, such as missing entries and imbalanced outcomes and provide healthcare providers with individualised risk assessments, potentially enhancing its applicability beyond the study population.

To achieve this, we developed statistical diagnostic tools to evaluate the performance of our ML model by addressing three key challenges. First, we addressed the challenge of imbalanced outcome distribution, a scenario where instances of poor clinical endpoints (ie, HIV care disengagement, high viral load or death) are rare, as failing to account for this imbalance could lead to biased predictions. Second, we examined the heterogeneity in the algorithm generated from multisite data to understand whether tailored, site-specific algorithms may have better performance than a single universal algorithm. Third, we evaluated the model's sensitivity to changes in virological suppression thresholds, analysing how changing the WHO-recommended cut-off from 1000 copies/mL to lower thresholds, such as 50 copies/mL, impacts the accuracy of risk score and informs future model adjustments.

Study sample

Using data from the national EMR repository including digitised medical records manually entered postclinical visits and managed by the Tanzania Ministry of Health (MoH) and the National AIDS, STIs and Hepatitis Control Programme, the original database included over 6 million records describing HIV primary care visits across Geita and Kagera regions in Lake Zone, Tanzania. This database, covering January 2015 to May 2023, described a cohort of 295,961 PLHIV and included 78 variables related to demographic profiles, clinical indicators and pharmaceutical records.

Outcome variables and candidate predictors

Aiming to guide the targeted implementation of proactive interventions for PLHIV to improve retention in HIV care, reduce morbidity and mortality and prevent onward transmission, we developed an algorithm targeting three key outcomes: care disengagement (at least one instance of missing a scheduled clinical visit by 28 days or more); high viraemia, (viral load greater than 1000 copies/mL at least once in their clinical visit history) and death.

To ensure the practicality of the proposed ML algorithm, we employed a bidirectional, participatory approach for

predictor selection among 78 covariates, collaborating with local clinicians and MoH representatives included in the HIV care EMR dataset. This process is detailed comprehensively in online supplemental appendix 1. We selected 16 covariates as predictors, categorised as either 'static' or 'dynamic', based on their characteristics throughout the study period. Static predictors, such as sex and initial referral location, remained constant for an individual. Conversely, dynamic predictors, including weight, ART status and viral load, vary from one visit to another, offering real-time insights into the patient's evolving status. For instance, ART status denotes an individual's current treatment status regarding ART, which may vary across different visits. During each visit, an individual may either initiate, change, continue, or stop ART, or not have initiated ART yet.

To capture the time-varying trend of dynamic predictors, we employed a regression approach to model their linear and quadratic changes over time, using derived quantities as predictors in our model and replacing the original dynamic variables. Consider the variable ART status as an example. Each individual might have multiple visit records, and the measurements of ART status can vary across these visits, making direct application of ART status to the algorithm challenging. To address this, we run a regression model with ART status as the dependent variable and the visits as the independent variable. This allows us to derive surrogate quantities that can be directly used in the algorithm. Specifically, we model the relationship between ART status and visits with the following equation:

$$ART = ART\ intercept + ART\ linear\ change * visit + ART\ quadratic\ change * visit^2$$

After running the regression, we use the derived quantities—ART intercept, ART linear change and ART quadratic change—as predictors in the algorithm. To enhance the identification of individuals at risk of loss to follow-up (LTFU), our model harnessed historical covariate information to forecast patients' future risk profiles, enabling more accurate and timely identification of individuals susceptible to potential LTFU.

ML algorithm development

We developed an ensemble decision trees algorithm with a downsampling strategy tailored to the unique challenges of the HIV care data.[23] This algorithm outperforms traditional ML approaches in two key aspects:

1. We employed the 'surrogate splits' technique in decision tree construction to handle missing data. Here, a surrogate split variable is identified to closely approximate the optimal split variable at nodes with missing values. This approach manages high levels of missingness in the HIV care data, ensuring incomplete observations are included in training, thus increasing the prediction power.
2. Traditional ML algorithms often produce biased results due to significant outcome imbalances in HIV care data, particularly in predicting the minority class, which is often the primary interest. To mitigate this bias, we employed downsampling, drawing a bootstrap sample from the minority class and a subsample from the majority class with matching sizes and combining them to build the decision tree. This strategy ensured a more equitable representation of both classes in model training, improving the model's ability to discern patterns associated with the minority class.

Moreover, to improve the accuracy and robustness, we used the ensemble decision tree model by conducting 1000 replications of downsampling strategy to build 1000 decision trees, aggregating their predictions for the final output. While downsampling can potentially lead to information loss from the majority class, our use of the ensemble approach mitigates this by drawing different subsamples in each iteration, effectively using the entire dataset's information over multiple trees.

With this algorithm, we computed risk scores for each individual and identified those at high risk of disengagement, high viraemia and death. By setting various risk thresholds, we can determine the proportion of the highest-risk individuals. To comprehensively assess model performance and guide the selection of practical risk thresholds, we split the data as shown in online supplemental appendix 2A, with the dataset from January 2018 to December 2021 serving as the training sample, and January 2018 to May 2023 as the testing sample. Subsequently, we built and validated the future outcomes model based on past predictors, evaluating sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV) across different thresholds.

Statistical diagnostic of the ML algorithm

Comparison with ensemble decision tree without downsampling

To evaluate the effectiveness of the downsampling strategy in handling imbalanced outcome distribution, we compared our algorithm with ensemble decision tree models with and without downsampling, focusing on the area under the curve (AUC) performance.

Comparison between site-specific training and all available data training

The HIV care data exhibited significant heterogeneity. To investigate the impact of this heterogeneity on the algorithm, we compared a universal algorithm (Algorithm (a), trained on data from all 373 sites included in Tanzania's National HIV Care and Treatment database) with site-specific algorithms (algorithm (b), trained and tested on data at the combination of clinics A and B only). The decision to focus exclusively on clinics A and B was driven by the specific applied goals of this research: developing and implementing the ML algorithm in these two pilot sites to identify high-risk patients for a subsequent study of a supportive intervention (NCT05373095). Both algorithms were evaluated using data at the combination of clinics A and B, following the training/testing split shown in online supplemental appendix 2A. This comparison

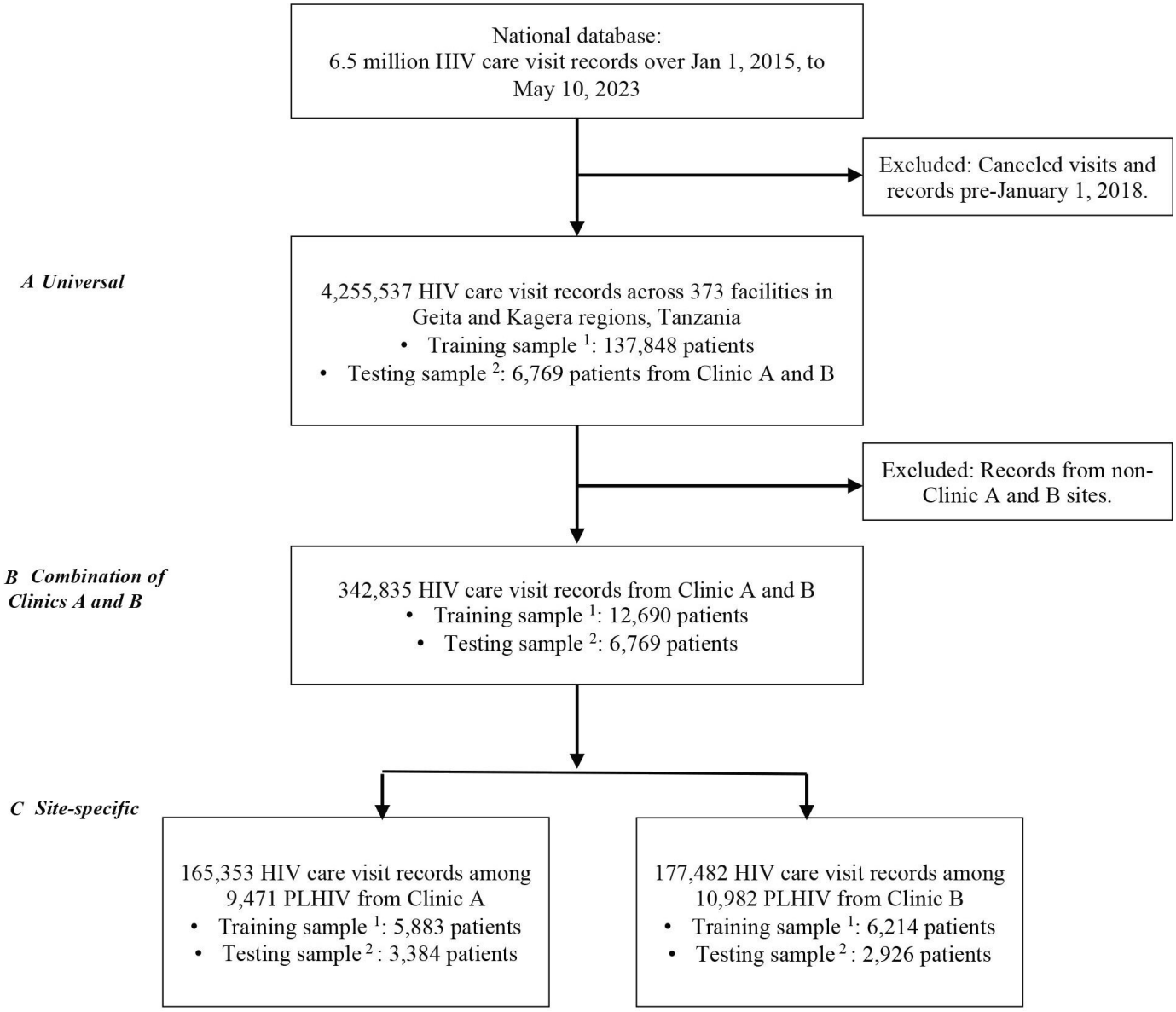

Figure 1 Flow diagram of the analysis population of PLHIV from Geita and Kagera regions, Tanzania (2018–2023).
[1]Training sample time frame: 1 January 2018–31 December 2021.
[2]Testing sample time frame: 1 January 2018–10 May 2023. PLHIV, people living with HIV.

aimed to examine how clinical site heterogeneity affects the algorithm's ability to predict high-risk individuals, evaluated by AUC, sensitivity, specificity, PPV and NPV across different risk thresholds.

On development of algorithm (b), we observed that the predicted high-risk individuals were skewed towards a particular clinical site. Thus, we proposed a new measure of future prediction fairness (FPF), defined as:

$$FPF = \left| \frac{\text{The number of High risk individuals in Clinic A}}{\text{The number of High risk individuals in Clinic A and Clinic B}} - \frac{\text{The number of High risk individuals in Clinic B}}{\text{The number of High risk individuals in Clinic A and Clinic B}} \right|$$

For equitable treatment across sites, an ideal algorithm should have a low FPF score. To further diagnose the potential disparity of the algorithm, we compared algorithm (b) with an alternative algorithm (algorithm (c)) that could achieve a low FPF. For evaluating FPF, we followed the training/testing data split procedure outlined in online supplemental appendix 2B.

In algorithm (c), we trained and applied the model on clinics A and B separately, which yielded two lists of risk scores for each facility. By setting different risk score thresholds for clinics A and B, we aimed to equalise the proportions of high-risk individuals among all individuals in both facilities. We evaluated both algorithms on AUC, sensitivity/specificity/PPV/NPV and FPF.

Sensitivity analysis with changing viral load threshold

We conducted a sensitivity analysis to assess the impact of changing thresholds for viral suppression on predicted high-risk individual profiles by modifying the viral load cut-off from the conventional WHO-recommended 1000 copies/mL, as used in the primary algorithm, to a more

Table 1 Dataset description; characteristics of PLHIV at their first visits since January 2018 from Geita and Kagera regions, Tanzania

	Clinic A	Clinic B	All sites (n=373)
Static variables	N=9471 patients	N=10 982 patients	N=253 208 patients
Age in years (median, IQR)	36.3 (17.1)	35.0 (15.7)	35.6 (16.4)
Sex (%)			
Male	3253 (34.3%)	3606 (32.8%)	96 691 (38.2%)
Female	6218 (65.7%)	7376 (67.2%)	156 517 (61.8%)
Time from HIV diagnosis to ART start, days (median, IQR)	724 (2086)	464 (1171)	392 (1053)
Time on ART, days (median, IQR)	757 (1887)	512 (1191)	464 (1086)
High viral load* (mean, SD)	0.6% (3.5%)	0.5% (2.5%)	0.5% (2.6%)
Disengagement† (mean, SD)	6.0% (11.3%)	9.1% (11.6%)	7.3% (14.6%)
Dynamic variables	N=165 353 visits	N=177 482 visits	N=4 255 537 visits
ART status‡ (%)			
Not started ART	25 (0.02%)	27 (0.02%)	193 (0.005%)
Start ART	6768 (4.1%)	6416 (3.6%)	170 296 (4.0%)
Continue ART	151 959 (91.9%)	163 316 (92.0%)	3 909 752 (91.9%)
Change ART	6563 (4.0%)	7694 (4.3%)	174 427 (4.1%)
Restart ART	31 (0.02%)	4 (0.002%)	341 (0.008%)
Stop ART	5 (0.003%)	4 (0.002%)	381 (0.009%)
Missing values	2 (0.001%)	21 (0.01%)	147 (0.003%)

Not started ART: Entries where ART has not been initiated. Start ART: Entries documenting the initiation of ART. Continue ART: Entries where patients continue with their ART without changes. Change ART: Entries reflecting a switch in ART regimen/plan. Restart ART: Entries for patients resuming ART after a previous discontinuation. Stop ART: Entries documenting the stop of ART. Missing values: Entries where the ART status is not recorded or is unknown.
*High viral load: percentage of instances where patients recorded a high viral load among all their visits.
†Disengagement: the proportion of instances in which patients were recorded as disengaged from ART among all their visits.
‡ART status: the percentage distribution of recorded ART statuses, reflecting changes or continuities in ART.
ART, antiretroviral therapy; PLHIV, people living with HIV.

stringent measure of 50 copies/mL, in line with upcoming guideline changes in Tanzania and input from MoH partners. This analysis was designed to observe shifts in patient classification as high-risk, potentially refining our understanding of the importance of viral load as a predictor and its correlation with the risk of adverse outcomes.

RESULTS

Descriptive data

The initial dataset contained records from 295 961 PLHIV, refined for algorithm use by focusing on completed and scheduled visits on or after 1 January 2018, resulting in visit information from 253 208 PLHIV across clinics A and B. Detailed data processing steps are illustrated in figure 1. Demographic and clinical information such as age, gender, referral location and ART status is summarised in table 1. The median age of PLHIV was approximately 35 years, with clinic B's cohort being slightly younger than clinic A. Typically, clinic A patients started ART 72 days postinitial visit while clinic B's median start time was 46 days.

Description of universal algorithm

In the initial algorithm, we employed data from all 373 facilities included in the HIV care database to train a universal algorithm. Following the split procedure shown in online supplemental appendix 2A, our training sample comprised 137 848 PLHIV from 2018 to 2021 across all sites. The testing sample included 6769 patients from January 2022 to May 2023 across clinic A and clinic B. In the assessment across all sites, the universal algorithm's efficacy varied by risk threshold, with an AUC of 0.583 (95% CI, 0.564 to 0.602) (figure 2A). At a 10% threshold defining the risk group, the universal algorithm demonstrated a sensitivity of 0.19 and a specificity of 0.91 (online supplemental appendix 3A).

Identification of key risk factors

Among all 16 predictors, the most important predictors associated with the composite outcome (disengagement, high viraemia or death) included past disengagement (static), ART status (dynamic), mean age during past visits (static), mean time on ART during past visits (static). The importance of these variables is detailed in figure 3.

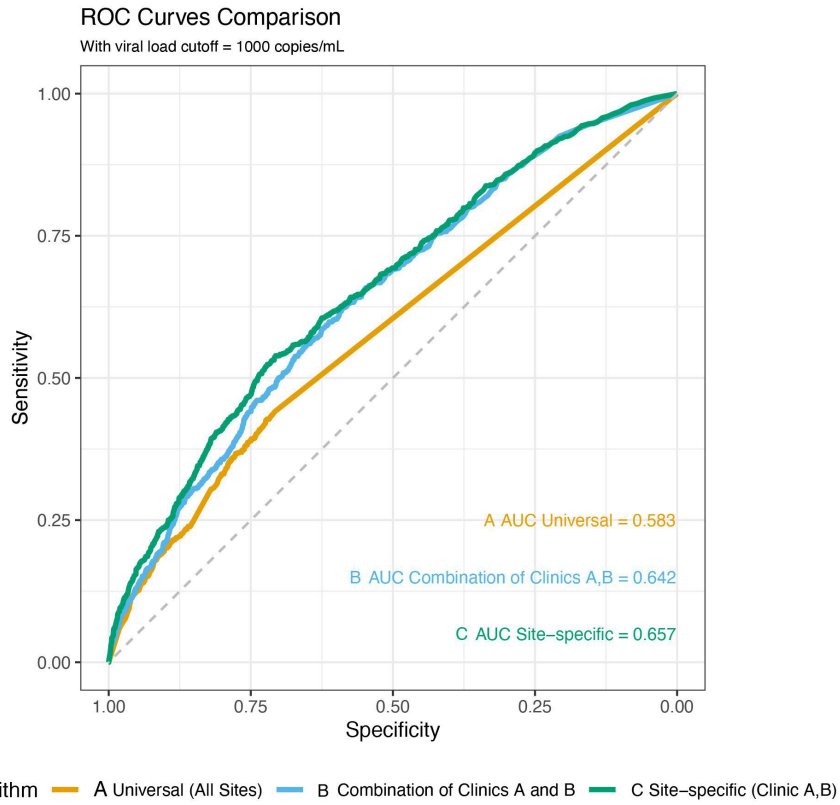

Figure 2 Comparative ROC curves for universal and site-specific algorithms: The universal algorithm (yellow) is evaluated using data from 137 848 patients across 373 sites in Geita and Kagera, Tanzania (2018–2021) and applied to 6679 patients (2018–2023.5) with a viral load threshold of 1000 copies/mL. The algorithm (blue) is assessed on 12 690 patients (2018–2021) and tested on 6769 patients (2018–2023.5) at clinic A and clinic B, using the same viral load threshold. The algorithm (green) is trained on 5883 clinic A patients (2018–2021) and applied to 3384 clinic A patients (2018–2023.5) and trained on 6214 clinic B patients (2018–2021) and applied to 2926 clinic B patients (2018–2023) separately, using the same viral load threshold. ROC, receiver operating characteristic.

Statistical diagnostic results for the ML algorithm

Imbalanced outcomes

In examining the impact of the downsampling strategy, we found that the algorithm without downsampling failed to provide useful prediction, predicting all individuals in the test sample with equal risk scores. Thus, all individuals in the test sample have equal probabilities being predicted as high risk, regardless of the risk threshold selection. This results in an AUC equal to 0.5 for both universal and site-combined (clinics A and B) models. Conversely, the algorithm that employed the downsampling strategy enhanced the algorithm performance with an AUC equal to 0.583 (95% CI 0.564 to 0.602) for the universal algorithm and an AUC of 0.642 (95% CI 0.622 to 0.662) for the algorithm using data at the combination of clinic A and clinic B only, effectively reducing bias in predictions.

Universal versus site-specific algorithms

The data split for training and testing of Algorithm (b), as detailed in figure 1 and online supplemental figure S2A, included 12 690 PLHIV for training and 6769 PLHIV for testing, from both clinics A and B. Algorithm (b) demonstrated better performance with an AUC of 0.642 (95% CI 0.622 to 0.662), compared with 0.583 (95% CI 0.564 to 0.602) for the algorithm (a) (figure 2A,B). Moreover, algorithm (b) showed enhanced sensitivity and PPV. For example, at the top 15% risk threshold, the PPV of the algorithm increased from 18.9% to 22.4% and the sensitivity increased from 23.6% to 28.0% compared with algorithm (a).

Further comparison between algorithms (b) and (c) involved evaluating AUC, sensitivity, specificity, PPV, NPV and FPF under various risk thresholds (figure 2). We found the sensitivity and PPV showed consistent improvement for the algorithm (c) (online supplemental appendix 3C) compared with the algorithm (b) (online supplemental appendix 3B). For example, at the top 10% risk threshold, PPV increased from 22.9% to 27.3% and the sensitivity increased from 19.0% to 21.6%, with AUC of 0.657 (95% CI 0.637 to 0.678) improving on algorithm (b).

Regarding FPF, we followed a different data split procedure as shown in online supplemental appendix 2B, including 3384 PLHIV from Clinic A and 2926 PLHIV from clinic B in a training sample to predict future high-risk individuals. We found that algorithm (b) skewed high-risk predictions towards patients from clinic A. For the top 7% risk scores, 882 out of 886 individuals flagged

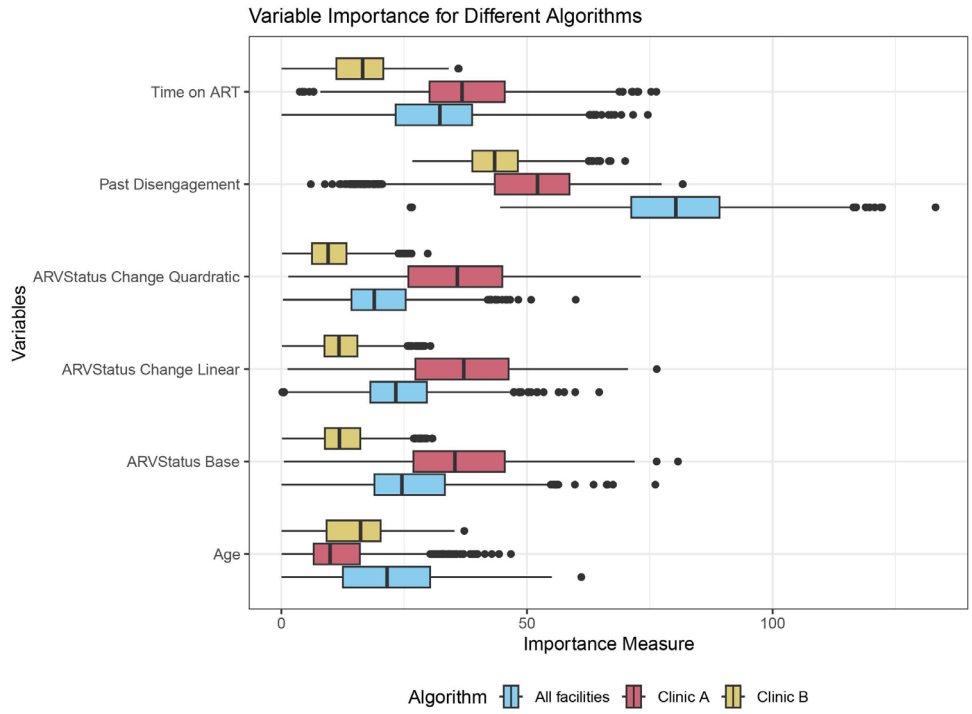

Figure 3 Variable importance in algorithms predicting outcomes for PLHIV. This figure illustrates the variable importance distribution in the prediction of patient outcomes using three different algorithms, based on HIV care data from PLHIV in the Geita and Kagera regions of Tanzania, spanning from 2018 to 2023. PLHIV, people living with HIV.

as high risk were from clinic A, resulting in an FPF of 0.991. High-risk predictions by algorithm (c) were more balanced, with 414 individuals from clinic A and 419 individuals from clinic B, with an FPF of 0.006. Thus algorithm (c) demonstrated an ability to enhance equity and the representativeness of the high-risk patient individuals generated by the algorithm in clinics A and B.

Changing viral load threshold

After adjusting the viral load cut-off in our algorithm from 1000 to 50 copies/mL, we observed consistent identification of high-risk individuals. Specifically, within the top 7% highest risk category, 82% of the same patients were classified as high risk, even with the stricter viral suppression definition. This indicated a strong correlation between risk scores and patient profiles, demonstrating the robustness of the algorithm across different viral load thresholds.

DISCUSSION

Our work applying ML models to digitised HIV care and treatment data from PLHIV in Tanzania demonstrates the technology's potential to enhance patient care through predictive insights, despite facing challenges common to EMR data like data quality and the need for custom algorithms suited to diverse healthcare environments. This research offers insights into addressing these challenges while leveraging the significant opportunities ML and EMR data present for advancing HIV healthcare in low-middle resource settings.

In our study, downsampling addressed the issue of imbalanced outcomes, a common challenge when dealing with rate outcomes in healthcare data. By balancing the data, downsampling enhanced our ML model's capacity to accurately identify at-risk individuals, thereby informing more targeted interventions within HIV care. While our focus was on improving the prediction of patient disengagement and high viraemia, the efficacy of this approach underscores its potential applicability across various clinical datasets beyond HIV care and can enhance the prediction of infrequent but important outcomes.

An important finding of our study is that given heterogeneity across clinical sites, a one-size-fits-all ML algorithm was suboptimal. This finding is broadly applicable across many health domains and healthcare settings, especially when resources are limited. This heterogeneity drove us to use a tailored dataset and algorithm for the two health facilities, which revealed an imbalance in high-risk patient identification, emphasising the need for site-specific model adjustments. The importance of incorporating site-specific details into our ML models has become evident, underscoring that understanding and integrating these local nuances is crucial for the predictive accuracy and utility of such models in healthcare. Acknowledging and addressing this heterogeneity can lead to a more effective allocation of resources and tailored interventions.[24]

Earlier studies in low-income and lower-middle-income countries have shown that poor adherence to ART is mainly associated with factors such as sex, age, concurrent

medications and care accessed at public healthcare facilities.[25] Our algorithm identified past disengagement, ART status, age and time on ART as significant predictors at both clinics. This information may be important for improving organisational decision-making.[26] Notably, while viral load testing is beneficial and viral load value is a critical metric on an individual level, previous studies and our model have not identified it as a paramount predictor at the model level.[27] Our sensitivity analysis that assessed a viral load threshold at a more stringent cut point showed that 82% of the patients remained classified as high risk. This stability suggests that while the viral load is a critical component of the risk model, the holistic patient profile, which encompasses a multitude of clinical and demographic variables, predominantly drives the risk predictions.

Our study's key strength is the use of a national EMR database, demonstrating the power of ML models to pinpoint patients most at risk of HIV care disengagement in resource-limited settings. Despite inherent challenges with large-scale EMR data, the insights gleaned are invaluable, and applicable beyond HIV care into border areas such as maternal and children's health. Furthermore, our participatory data science approach, involving collaboration among clinicians, academics and policy-makers, significantly enriches our research and highlights the importance of engaging stakeholders in practical ML applications.[28]

One notable limitation of this study is the reliance on periodic downloads of the HIV care database rather than real-time data access, potentially introducing challenges in future applications due to data lag. In addition, the HIV care database recorded all-cause mortality instead of cause-specific mortality, limiting the information available on HIV-related deaths. Furthermore, the reliability of EMR data presents inherent limitations. Although we incorporated methods to address data missingness, as with any EMR data, incompleteness or inaccuracies in the data may affect the performance of ML models.

Conclusion

This study demonstrated the significant potential of our ML model to identify PLHIV at increased risk of adverse outcomes, successfully navigating issues of imbalanced outcomes, heterogeneity across clinical sites and evolving clinical guidelines. By deploying the algorithm using digitised health record data, healthcare providers can improve the targeting and implementation of interventions and ultimately advancements towards UNAIDS' 2030 targets. Our research not only illuminates pathways for enhancing engagement and retention in HIV care but also serves as a testament to the power of innovative data science in public health strategies in resource-limited settings.

Contributors All coauthors conceptualised the study. ZX, SIM and JW developed the methodology. ZX, JLK, LJP and MM contributed to data curation. ZX and HH wrote the original draft of the report. ZX, HH, JLK and JW were involved in the investigation, formal analysis, validation and visualisation. JLK, SK, WM, PFN and SS had full access to raw data. All coauthors contributed to the review and editing of the report. AS, SIM and JW contributed to the supervision of the study and were responsible for funding and computational and data analytic resources. ZX, HH, SIM and JW had final responsibility for the overall content and the decision to submit for publication as guarantor. The corresponding author confirms that all authors have seen and approved the final manuscript.

Funding This study was supported by a grant from the US National Institutes of Health (NIH): NIH NIMH 1R01MH125746 (SM).

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval The Tanzania National Institute for Medical Research (NIMR) and the Committee for Protection of Human Subjects at the University of California, Berkeley (UC Berkeley) provided ethical review and approval for this study. (UC Berkeley ethics approval number: 2021-10-14723, obtained on 17 December 2021; Tanzania NIMR ethics approval number: NIMR/HQ/R.8a/Vol. IX/3992, obtained on 5 May 2022). With approval from both ethical review boards, we received a waiver of written informed consent for this part of study in accordance with FDA guidelines according to 45 CFR 46.116(f). Specific permission was granted to use routinely collected demographic and clinical information, as well as the minimal amount of identifiable private information to identify participants for potential future trials.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. The data used in this study are owned by National AIDS, STIs and Hepatitis Control Programme (NASHCoP) and Ministry of Health (MoH), Tanzania. The data access is restricted to authorised researchers under data use agreements to ensure confidentiality and compliance with the Protection of Personal Information Act (POPIA); although data cannot be shared publicly, interested parties can request data and propose its use to the authorship team, who can present it to NASHCoP/MoH for consideration.

ORCID iDs
Huiyu Hu http://orcid.org/0009-0000-0182-6654
Sandra I McCoy http://orcid.org/0000-0002-4764-9195

REFERENCES

1 UNAIDS. *Understanding fast-track: accelerating action to end the AIDS epidemic by 2030*. 2015.
2 Lifson AR, Grund B, Gardner EM, *et al*. Improved quality of life with immediate versus deferred initiation of antiretroviral therapy in early asymptomatic HIV infection. *AIDS* 2017;31:953–63.
3 Cohen MS, Chen YQ, McCauley M, *et al*. Antiretroviral Therapy for the Prevention of HIV-1 Transmission. *N Engl J Med* 2016;375:830–9.
4 Fox MP, Rosen S. Retention of Adult Patients on Antiretroviral Therapy in Low- and Middle-Income Countries: Systematic Review and Meta-analysis 2008-2013. *J Acquir Immune Defic Syndr* 2015;69:98–108.
5 World Health Organization. *Retention in HIV programmes: defining the challenges and identifying solutions: meeting report, 13-15 September 2011*. World Health Organization, 2012.
6 Bonner K, Mezochow A, Roberts T, *et al*. Viral Load Monitoring as a Tool to Reinforce Adherence. *J Acquir Immune Defic Syndr* 2013;64:74–8.

7 Glasner S, Chokron Garneau H, Ang A, *et al*. Preliminary Efficacy of a Cognitive Behavioral Therapy Text Messaging Intervention Targeting Alcohol Use and Antiretroviral Therapy Adherence: A Randomized Clinical Trial. *PLoS One* 2020;15:e0229557.
8 Fahey CA, Njau PF, Katabaro E, *et al*. Financial incentives to promote retention in care and viral suppression in adults with HIV initiating antiretroviral therapy in Tanzania: a three-arm randomised controlled trial. *Lancet HIV* 2020;7:e762–71.
9 Clouse K, Noholoza S, Madwayi S, *et al*. The Implementation of a GPS-Based Location-Tracking Smartphone App in South Africa to Improve Engagement in HIV Care: Randomized Controlled Trial. *JMIR Mhealth Uhealth* 2023;11:e44945.
10 Fatti G, Ngorima-Mabhena N, Mothibi E, *et al*. Outcomes of Three-Versus Six-Monthly Dispensing of Antiretroviral Treatment (ART) for Stable HIV Patients in Community ART Refill Groups: A Cluster-Randomized Trial in Zimbabwe. *J Acquir Immune Defic Syndr* 2020;84:162–72.
11 Maskew M, Technau K, Davies M-A, *et al*. Adolescent retention in HIV care within differentiated service-delivery models in sub-Saharan Africa. *Lancet HIV* 2022;9:e726–34.
12 Fraser HSF, Mugisha M, Remera E, *et al*. User Perceptions and Use of an Enhanced Electronic Health Record in Rwanda With and Without Clinical Alerts: Cross-sectional Survey. *JMIR Med Inform* 2022;10:e32305.
13 Mashoka RJ, Murray B, George U, *et al*. Implementation of electronic medical records at an Emergency Medicine Department in Tanzania: The information technology perspective. *Afr J Emerg Med* 2019;9:165–71.
14 Mamo DN, Yilma TM, Fekadie M, *et al*. Machine learning to predict virological failure among HIV patients on antiretroviral therapy in the University of Gondar Comprehensive and Specialized Hospital, in Amhara Region, Ethiopia, 2022. *BMC Med Inform Decis Mak* 2022;23:75.
15 Pence BW, Bengtson AM, Boswell S, *et al*. Who Will Show? Predicting Missed Visits Among Patients in Routine HIV Primary Care in the United States. *AIDS Behav* 2019;23:418–26.
16 Maskew M, Sharpey-Schafer K, De Voux L, *et al*. Applying machine learning and predictive modeling to retention and viral suppression in South African HIV treatment cohorts. *Sci Rep* 2022;12:12715.
17 Obermeyer Z, Emanuel EJ. Predicting the Future - Big Data, Machine Learning, and Clinical Medicine. *N Engl J Med* 2016;375:1216–9.
18 Fahey CA, Wei L, Njau PF, *et al*. Machine learning with routine electronic medical record data to identify people at high risk of disengagement from HIV care in Tanzania. *PLOS Glob Public Health* 2022;2:e0000720.
19 Fu S, Leung LY, Raulli A-O, *et al*. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. *BMC Med Inform Decis Mak* 2020;20:60.
20 Schwartz JT, Gao M, Geng EA, *et al*. Applications of Machine Learning Using Electronic Medical Records in Spine Surgery. *Neurospine* 2019;16:643–53.
21 Sauer CM, Chen L-C, Hyland SL, *et al*. Leveraging electronic health records for data science: common pitfalls and how to avoid them. *Lancet Digit Health* 2022;4:e893–8.
22 Ellman TM, Alemayehu B, Abrams EJ, *et al*. Selecting a viral load threshold for routine monitoring in resource-limited settings: optimizing individual health and population impact. *J Int AIDS Soc* 2017;20 Suppl 7:e25007.
23 Branco P, Torgo L, Ribeiro R. A Survey of Predictive Modelling under Imbalanced Distributions. 2015.
24 Sculpher M. Subgroups and heterogeneity in cost-effectiveness analysis. *Pharmacoeconomics* 2008;26:799–806.
25 Bijker R, Jiamsakul A, Kityo C, *et al*. Adherence to antiretroviral therapy for HIV in sub-Saharan Africa and Asia: a comparative analysis of two regional cohorts. *J Int AIDS Soc* 2017;20:21218.
26 Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLoS One* 2019;14:e0212356.
27 Drain PK, Dorward J, Bender A, *et al*. Point-of-Care HIV Viral Load Testing: an Essential Tool for a Sustainable Global HIV/AIDS Response. *Clin Microbiol Rev* 2019;32:e00097-18.
28 Qiao S, Li X, Olatosi B, *et al*. Utilizing Big Data analytics and electronic health record data in HIV prevention, treatment, and care research: a literature review. *AIDS Care* 2024;36:583–603.