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.

A Multi Facility Data Quality Assessment of Electronic Health Records for Trustworthy AI in HIV Public Health in Zambia

Joe Phiri

phirijoe26@hotmail.com

ZCAS University

Aaron Zimba

ZCAS University

Chiyaba Njovu

ZCAS University

Andrew Kashoka

Ministry of Health

Innocent Chiboma

Ministry of Health

Mulenga Chiwele

Ministry of Health

Trevor Sinkala

Centre for Infectious Disease Research in Zambia (CIDRZ)

Jacob Mutale

Centre for Infectious Disease Research in Zambia (CIDRZ)

Mwansa Lumpa

Centre for Infectious Disease Research in Zambia (CIDRZ)

Research Article

Keywords: Electronic health records, Data quality, HIV, Artificial intelligence, Machine learning, Public health, Algorithmic fairness, SHAP, Sub-Saharan Africa

Posted Date: May 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9306862/v1

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

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Discover Artificial Intelligence on September 21st, 2026. See the published version at https://doi.org/10.1007/s44163-026-02293-x.

Abstract

Background

Artificial intelligence (AI) holds promise for HIV disease surveillance and clinical decision support in sub-Saharan Africa. However, AI reliability and fairness depend fundamentally on data quality, a dimension rarely assessed before model development. This study presents a systematic, multi-facility data quality assessment (DQA) of Zambia's national HIV electronic health record (EHR) system, using an active AI/machine learning (ML) pipeline as the diagnostic lens.

Methods

We applied an eight-dimension DQA framework: completeness, conformance, plausibility, feature coverage, outcome sensitivity, temporal stability, cross-facility concordance, and missingness-outcome confounding to 246,053 patient records across six Lusaka public health facilities. SHAP attribution patterns from a trained AI model surfaced latent data quality problems.

Results

Education level completeness was 1.2–7.9% across all facilities, household income 4.7–19.8%, and HIV enrolment stage 21.1–24.6%. Engagement features were computable for only 12.2% of patients; laboratory features for 3.3%. Among patients with engagement data, 54.3% had appointment lateness values exceeding plausible limits (maximum: 45,459 days). Facility-level missingness rates correlated strongly with outcome rates — education vs. viral suppression ($r = -0.865$), HIV stage vs. tuberculosis incidence ($r = +0.741$) — indicating the AI model partly learned data entry patterns rather than clinical risk. A tuberculosis outcome prevalence of 36% reflects label conflation of prevalent and incident TB cases. The composite Data Quality Index ranged from 53.5% to 58.3%, all below the 70% threshold.

Conclusions

Sociodemographic data gaps create an algorithmic fairness risk: AI models may flag patients for missing records rather than genuine clinical need. We propose eight minimum data quality requirements and a reproducible DQA toolkit for trustworthy, equitable AI in resource-limited settings, aligned with SDG 9 and SDG 10.

Introduction

Artificial intelligence (AI) and machine learning (ML) are increasingly applied to public health challenges in low- and middle-income countries (LMICs). In HIV care, AI systems trained on routine health records have shown promising results for predicting which patients are at risk of treatment failure, tuberculosis

co-infection, and loss to follow-up [1], [2]. These tools could help health workers and programme managers prioritise attention and resources in settings where staff and budgets are severely constrained. However, the promise of AI in public health is only as strong as the data it learns from.

The quality of routine electronic health records (EHRs) in resource-limited settings is fundamentally different from the high-income healthcare contexts where most AI data quality research originates [3], [4]. In settings like Zambia, EHR data entry is often performed retrospectively by data clerks working from paper records, computer literacy is variable, internet connectivity is unreliable, and the same information system is shared across dozens of health facilities that each operate slightly differently. When AI is trained on data with these characteristics, it may learn patterns in who has missing data rather than patterns in who is at clinical risk.

This distinction matters for public health equity. If an AI system flags patients for health interventions primarily because their demographic records are incomplete rather than because they have genuine health needs, it risks misallocating resources and potentially stigmatising patients whose records are missing for administrative rather than clinical reasons. This directly conflicts with the goals of SDG 10 (Reduced Inequalities) and the principle of ethical AI in public health [5].

This study addresses this gap through a systematic, empirically grounded data quality assessment of the SmartCare EHR system — Zambia's national HIV electronic records platform using data from 246,053 patients across six Lusaka public health facilities. We use an active AI pipeline as our diagnostic instrument: the model's feature importance rankings and SHAP (Shapley Additive Explanations) attribution patterns reveal data quality problems that standard statistical checks would miss. Our approach is directly relevant to the collection's focus on ethical AI, social determinants of health, and community-based data integration for equity-focused public health solutions.

The study is motivated by a striking finding from a companion AI study [6]: education level, a field with 1.2–7.9% completeness across all six facilities became the single most important AI predictor of HIV viral load suppression failure, contributing up to + 2.58 units of log-odds risk for individual patients. This finding raises important questions: Is this because patients without recorded education genuinely have lower health literacy and poorer treatment outcomes? Or is the AI learning that certain facility types which serve higher-risk populations and also have worse data entry practices, tend not to record patients' education levels? These two explanations have completely different policy implications, and the data quality assessment is what makes it possible to begin separating them.

We make five contributions: (i) the systematic multi-facility DQA of Zambia's national HIV EHR; (ii) an eight-dimension DQA framework for AI applications in resource-limited HIV programmes; (iii) quantification of feature coverage collapse, a form of AI training data sparsity that is rarely reported; (iv) a formal missingness-outcome confounding test using SHAP attribution patterns; and (v) eight minimum data quality requirements for trustworthy AI in this context, with a reproducible toolkit embedded in the analytical pipeline.

Methods

2.1 Data source

We used data from 246,053 unique, deduplicated HIV patients across six public health facilities in Lusaka District, Zambia: Chawama First Level Hospital, Chelstone Urban Health Centre (UHC), George UHC, Kanyama First Level Hospital, Matero Main UHC, and Matero Reference Hospital. Four facilities were used to train the AI model (training set, n = 171,547) and two were held out for evaluation (test set, n = 74,506). Data were structured as four file types per facility: file1 — patient demographics and enrolment details; file2 — longitudinal visit and clinical event records; file3 — programme status and outcome records; and file4 — laboratory test results. The complete dataset comprised 24 CSV files.

This study used de-identified secondary programme data from the SmartCare EHR system. All methods were carried out in accordance with relevant guidelines and regulations, including the Declaration of Helsinki. Ethics approval was granted by the ZCAS University Ethics Review Committee (Reference No. 2025/11/001). Access to the data was authorised by the Ministry of Health, Zambia. Individual patient consent was not required.

Baseline characteristics of the full cohort, measured from raw pre-imputation data, are presented in Supplementary Table 8.

2.2 Eight-dimension data quality framework

We adapted the Weiskopf-Weng EHR data quality framework [7] to eight dimensions specific to AI applications in resource-limited HIV EHR systems. Table 1 defines each dimension and explains its role in ensuring the AI model is trained on reliable data.

Table 1
Eight-dimension data quality assessment framework for AI applications in resource-limited HIV EHR settings, adapted from Weiskopf and Weng [7]

ID	Dimension	Definition	Purpose
D1	Completeness	Proportion of records with non-null, non-'None' values per field, per facility.	Identifies fields and facilities with systematic recording gaps that force imputation.
D2	Conformance	Proportion of values conforming to expected data types, formats, and controlled vocabulary (e.g. WHO stage within {1,2,3,4}; dates parse-able without error).	Quantifies encoding inconsistencies: unit suffixes in numeric fields, non-standard date formats.
D3	Plausibility	Proportion of numeric values within clinically believable ranges: height 80–250 cm; weight 20–250 kg; appointment lateness within ± 730 days per visit.	Detects extreme errors such as lateness values of 45,459 days, that standard pipelines do not filter.
D4	Feature coverage	Proportion of patients for whom each derived feature group can be computed after restricting records to the period before the patient's enrolment date.	Measures the actual proportion of the training dataset with meaningful feature values, not just field presence.
D5	Outcome definition sensitivity	How much outcome prevalence changes across different facility subsets or operationalisation choices.	Reveals whether the AI outcome labels are stable enough to support cross-facility or cross-study comparisons.
D6	Temporal stability	Trends in field completeness over the years patients were enrolled (2003–2025).	Shows whether data quality has improved, worsened, or changed abruptly and when.
D7	Cross-facility concordance	Heterogeneity in field names (schema), value distributions, and missingness rates across the six facilities.	Measures how consistently data is recorded across facilities serving the same programme.
D8	ML missingness-outcome confounding	Pearson correlation between the facility-level rate of missing ('None') values and the facility-level outcome rate; supported by SHAP attribution analysis.	Tests whether AI predictions are driven by missing-data patterns rather than genuine patient risk.

2.3 Analysis procedures

For dimensions D1–D3 and D7, we computed per-field, per-facility completeness rates, value distribution summaries, and counts of implausible values across all 246,053 patients. A data point was considered 'complete' only if it was non-null, non-empty, and not coded as 'None'. Dimension D4 feature coverage was the proportion of patients who had at least one usable record in file2 or file4 after restricting to records before the patient's enrolment date. For D5, we computed outcome prevalence per facility and

for three data subsets: all facilities, training set, and test set. Dimension D6 stratified field completeness by patient enrolment year from 2003 to 2025. Dimension D8 computed the Pearson correlation between facility-level missing-data rates and facility-level outcome rates across the six facilities, and examined SHAP attribution patterns from the companion AI model [6].

A composite Data Quality Index (DQI) was computed for each facility as the unweighted average completeness rate across 15 AI-critical fields. We set 70% as a minimum quality threshold, below which mean imputation may distort model learning. All analyses were implemented in Python 3.13 (pandas, numpy, matplotlib, seaborn) as reproducible DQA framework.

2.4 AI pipeline as a diagnostic tool

The companion AI study [6] trained a stacked ensemble model (logistic regression, random forest, and XGBoost as base models; logistic regression as the combining layer) to predict three outcomes over 12 months: incident tuberculosis (TB12m), interruption in treatment (IIT12m), and unsuppressed viral load (UVL12m). Standard preprocessing was applied: median imputation for numeric features, most-frequent-value imputation for categorical features, standardisation, and one-hot encoding. All preprocessing was fitted on training data only. SHAP TreeExplainer [8] was applied to the XGBoost component on 800 held-out test patients per outcome. In this DQA study, the model's imputation volumes, permutation importance rankings, and SHAP attribution patterns are used as empirical evidence of underlying data quality issues.

Results

3.1 Dataset structure and cross-facility variation (D7)

The study cohort comprised 246,053 unique HIV patients (median age 32 years, IQR 26–38; 62.5% female). Among the 84.1% of patients with valid anthropometric data, median BMI was 21.8 kg/m² (IQR 19.1–25.2); 19.4% were underweight (BMI < 18.5 kg/m²). A total of 1,323 patients (0.5%) were on TB treatment at enrolment. Enrolments spanned 2003 to 2025, with 25.5% of patients enrolled before 2010 and 20.3% from 2020 onwards. Full pre-imputation baseline characteristics, including sociodemographic and clinical detail by facility, are presented in Supplementary Table 8.

Table 2 summarises the six facilities, their patient volumes, AI model roles, and DQI scores. The number of patients per facility ranged from 26,845 to 56,871 — a 2.1-fold difference reflecting different facility sizes and the length of time each has been recording data electronically. Importantly, all six facilities had below-70% DQI scores, confirming that the data quality problems identified here affect the programme as a whole, not individual outlier facilities.

The preprocessing code required 23 separate checks to handle cases where a data field was present in some facilities but missing or differently named in others. Each such check is a signal that at least one

facility records a given piece of information differently from the others, a form of cross-facility inconsistency that is invisible in aggregate statistics.

Table 2
Study dataset across 15 key AI features.

Facility	AI model role	Patients (N)	DQI score	Quality tier
KanyamaFirstLevelHospital	Training	56,871	58.3%	Highest DQI
ChawamaFirstLevelHospital	Training	47,314	57.3%	Good
MateroReferenceHospital	Test	47,661	55.8%	Moderate
ChelstoneUHC	Training	31,785	57.0%	Good
GeorgeUHC	Training	35,577	55.4%	Moderate
MateroMainUHC	Test	26,845	53.5%	Lowest DQI
All 6 facilities	—	**246,053**	**53.5–58.3%**	**All below 70% threshold**

3.2 Field completeness: a clear three-tier pattern (D1)

Table 3 and Fig. 1 show field completeness rates across the 15 AI-critical features for all six facilities. Three distinct tiers are evident.

High completeness (≥ 80%, green): Basic registration fields: gender, age, and TB treatment at enrolment are near-perfect (99–100%) at all facilities. Physical measurements (height 72–90%, weight 84–93%, BMI 70–88%) are generally high, though Matero Main UHC has the lowest BMI completeness at 70%.

Moderate completeness (40–79%, amber)

Marital status ranges from 70% to 80% across facilities incomplete but substantive. HIV enrolment stage is recorded for only 21–25% of patients. This is particularly important because HIV enrolment stage is the third most important feature for predicting TB incidence in the companion AI model, yet it is missing for three in every four patients.

Critical missingness (< 10%, red)

Education level is recorded for only 1.2–7.9% of patients. Household income is 4.7–19.8%. Past TB type is 0.5–1.3%. Exit reason and last interruption date are 0% at every facility. These are not fields with occasional missing values, they are fields that are effectively absent from the digital record system.

Table 3
Per-field, per-facility EHR completeness rates (%) for 15 key AI features.

Field	Chawama	Chelstone	George	Kanyama	MatRef	MatMain	Mean
Gender	99%	100%	100%	100%	100%	100%	100%
Age	99%	100%	100%	100%	100%	100%	100%
Marital status	78%	74%	70%	80%	70%	80%	75%
height	85%	89%	86%	89%	86%	72%	85%
Weight	89%	93%	88%	93%	88%	84%	89%
BMI	84%	88%	84%	88%	84%	70%	83%
TB_rx_enrolment	100%	100%	100%	100%	100%	100%	100%
HIV enrolment stage	24%	22%	25%	23%	21%	21%	23%
HIV enrolment status	18%	17%	13%	16%	15%	12%	15%
Education level	5%	5%	3%	8%	6%	1%	5%
Household income	20%	10%	5%	18%	7%	9%	11%
Past TB type	1%	1%	1%	1%	1%	1%	1%
Exit reason	0%	0%	0%	0%	0%	0%	0%
Last interruption date	0%	0%	0%	0%	0%	0%	0%

On the heat map: Columns = feature fields; rows = facilities. Colour = completeness rate (red = 0%; green = 100%). Education and household income appear as uniformly red columns, indicating near-total absence across all facilities. HIV enrolment stage is amber (partial recording). Basic registration fields (gender, age, tb_rx_enrolment) are uniformly green, confirming that poor completeness is field-specific, not a general system failure.

Notably, the near-absent fields identified here are education level and household income and are among the top predictors in the companion AI model [6]. For the unsuppressed viral load outcome, the missing-data indicator for education (‘None’) had the highest mean SHAP value of all 57 model features. The implications of this pattern for AI reliability and fairness are examined in Section 4.1.

3.3 Feature coverage collapse: the AI model's hidden training gap (D4)

Table 4 reveals the most consequential data quality finding. Despite 246,053 patients in the dataset, engagement features (appointment attendance patterns, visit history) were computable for only 30,065 patients (12.2%). Laboratory features (viral load history, CD4 counts, TB test results) were computable for only 8,233 (3.3%).

Table 4
Feature coverage

Feature type	Total patients	Patients with usable data	Coverage (%)	Zero-imputed (%)
Demographic (file1)	**246,053**	**246,053**	**100.0**	**0.0**
Engagement / visits (file2)	246,053	30,065	12.2	87.8
Laboratory results (file4)	246,053	8,233	3.3	96.7

This means that for 87.8% of patients, engagement-related features were derived from imputed or default values rather than observed data: visit count = 0, mean appointment lateness = imputed to training median, maximum appointment lateness = imputed to training median. For 96.7% of patients, all laboratory features were similarly imputed. The model was unable to learn meaningful clinical history signals from these patients, as predictions were largely driven by imputed or missing data patterns

Three mechanisms likely explain this collapse. First, patients enrolled in earlier years show substantially lower availability of engagement data, suggesting that historical records were not fully digitised or captured at the same level of detail as more recent data. Second, newly enrolled patients have limited or no prior visit history, resulting in missing engagement features and reducing the availability of meaningful input data for prediction. Third, data entry completeness for visit and laboratory records varies across facilities. The result is that the AI model was essentially trained on a one-time demographic snapshot for 87.8% of its training patients, making demographic indicators, including missing-data flags disproportionately influential.

3.4 Plausibility violations: errors that entered the AI model unchecked (D3)

Figure 2 shows the distribution of appointment lateness values among the 30,065 patients with engagement data, clipped to ± 365 days for display. Among these patients, 16,439 (54.3%) had maximum appointment lateness values exceeding 730 days which is impossible for a single appointment in a programme that has existed since 2004. The highest value recorded was 45,459 days (approximately 124 years). A further 3,965 patients (13.2%) had negative lateness values greater than 30 days, meaning the system recorded them attending more than a month before their scheduled appointment.

These extreme values entered the AI model without any plausibility check, since the current pipeline only applies range limits to height and weight. Appointment lateness is among the top-5 AI features for both TB and treatment interruption prediction, meaning the model's predictions for these outcomes are partly influenced by data entry errors rather than real clinical behaviour.

3.5 Outcome definition sensitivity: AI labels are not stable across facilities (D5)

Table 5 shows how the prevalence of each AI outcome variable changes across data subsets and facilities. The differences are important for understanding what the AI model is actually predicting.

Table 5
Outcome prevalence across data subsets and inter-facility variation

Outcome	Data subset	Positive cases (N)	Prevalence	Data quality note
TB12m	All facilities (N = 246,053)	88,585	36.0%	Facility range: 32.1%–48.4% — reflects real epidemiological differences
TB12m	Training set (N = 171,547)	57,812	33.7%	—
TB12m	Test set (N = 74,506)	30,773	41.3%	—
IIT12m	All facilities	1,378	0.56%	Facility range: 0.20%–0.65% — partly reflects inconsistent coding of inactivity flag
IIT12m	Training set	905	0.53%	—
IIT12m	Test set	473	0.63%	—
UVL12m	All facilities	3,564	1.45%	Facility range: 0.80%–2.19% — partly reflects differential viral load testing rates
UVL12m	Training set	2,693	1.57%	—
UVL12m	Test set	871	1.17%	—

Tuberculosis incidence (TB12m)

The 8.7-percentage-point difference between the training (33.7%) and test (41.3%) facility prevalence rates reflects genuine epidemiological differences, the test facilities serve populations with a higher TB burden. While this makes the AI modelling task harder, it is not a data quality problem.

Treatment interruption (IIT12m)

The overall prevalence is 0.56%, with a 3.3-fold range across facilities (0.20%–0.65%). This degree of variation for an outcome like treatment interruption, which should reflect broadly similar programme-

level patterns across Lusaka facilities suggests inconsistent coding of the programme inactivity flag in the records system. At some facilities, patients whose status was never updated appear incorrectly as inactive; at others, the flag is only set when confirmed disengagement is documented.

Unsuppressed viral load (UVL12m)

The 2.7-fold facility range (0.80%–2.19%) partly reflects the fact that facilities conducting more viral load tests per patient will detect more cases of unsuppressed viral load in the 12-month outcome window. A facility with low testing rates produces few detectable UVL events, this does not mean its patients are virologically better controlled.

These findings mean that published AI performance statistics for HIV treatment outcomes in sub-Saharan Africa cannot be meaningfully compared across studies unless the outcome definitions and facility mix are documented in the same way. This is a significant barrier to building a cumulative evidence base for AI in HIV public health.

3.6 Temporal stability: when in time was data quality better or worse? (D6)

Figure 3 shows field completeness stratified by patient enrolment year. The patterns reveal both progress and persistent problems.

Age completeness (pink) is 100% throughout. Marital status (green) was unusually low in 2004–2006 then recovered. HIV enrolment stage (red) was near-zero until 2016, peaked at approximately 70% in 2016–2019, then declined. Household income (orange) shows the same brief peak pattern. Education (blue) has never exceeded approximately 8% in any year, only rising slowly post-2022. Physical measurements (weight, height) have been consistently high throughout

Three important patterns stand out. First, basic registration data (age, gender, physical measurements) has been recorded reliably throughout the programme's history this confirms that poor completeness is not a fundamental system limitation but is field-specific. Second, HIV enrolment stage and household income completeness briefly improved in 2016–2019 before declining again, suggesting a data entry campaign or system update that was not sustained, a common pattern in resource-limited digital health implementation. Third, and most significantly, education completeness has never exceeded approximately 8% in any enrolment year over the entire period. Even as the SmartCare system has expanded and been upgraded, this field has remained effectively unused, which means the persistent missing-data flag for education is not just a historical legacy of old data, but an ongoing structural gap in the programme's data collection practices.

3.7 Missing-data patterns in AI predictions: the confounding test (D8)

Table 6 presents the core fairness and reliability test of this study: whether the AI model's predictions are partly driven by missing-data patterns rather than genuine patient risk.

Table 6
Missing-data confounding test.

Feature (missing-data indicator)	Outcome	Correlation	Interpretation
Education	TB12m	r = + 0.599	Likely confounding — facilities with more missing education records also have higher TB rates
Education	IIT12m	r = + 0.105	Low risk
Education	UVL12m	r = − 0.865	Likely confounding — facilities with less complete education data record fewer unsuppressed viral load events
Household income	TB12m	r = + 0.263	Low risk
Household income	IIT12m	r = + 0.064	Low risk
Household income	UVL12m	r = − 0.478	Borderline — investigate
Marital status	TB12m	r = − 0.429	Low risk
Marital status	IIT12m	r = − 0.028	Low risk
Marital status	UVL12m	r = − 0.181	Low risk
HIV enrolment stage	TB12m	r = + 0.741	Likely confounding — facilities with more missing stage records have higher TB rates
HIV enrolment stage	IIT12m	r = − 0.085	Low risk
HIV enrolment stage	UVL12m	r = + 0.126	Low risk

Three correlations exceed the likely-confounding threshold of |r| > 0.5. Understanding each one requires knowing the direction of the effect.

Education missing vs. TB incidence (r = + 0.599)

Facilities where more patients have no recorded education level also have higher TB12m rates. This is consistent with both genuine epidemiology (facilities serving more socially disadvantaged populations may have both lower health literacy and higher TB burdens) and data quality confounding (facilities with less complete data may also serve higher-risk catchment populations). The two explanations cannot currently be separated.

Education missing vs. viral load suppression (r = − 0.865): This is a strong negative correlation: facilities where more patients have no recorded education tend to have lower recorded rates of unsuppressed

viral load. This is counter-intuitive if the model were detecting genuine health literacy risk, it suggests that the AI may partly be learning that facilities with worse education recording also conduct fewer viral load tests (and therefore detect fewer unsuppressed cases). This is a data quality confound, not a clinical signal.

HIV enrolment stage missing vs. TB incidence (r = + 0.741)

Facilities where HIV enrolment stage is more often missing also have higher TB12m rates. This suggests the AI model may be detecting facility type, higher-burden facilities serving more complex patients may also have less complete clinical staging records, rather than individual patient risk.

These findings do not prove that the AI model's predictions are wrong. They demonstrate that the model cannot currently distinguish between a patient who is clinically at risk and a patient who attends a facility with poor data entry practices. For a public health AI tool to be equitable and trustworthy, this distinction must be resolvable, which requires either improving data quality or explicitly modelling and adjusting for the missing-data mechanism.

3.8 Composite Data Quality Index

Figure 4 shows the composite DQI per facility, the unweighted average completeness across 15 AI-critical fields. Scores range from 53.5% (Matero Main UHC) to 58.3% (Kanyama First Level Hospital). No facility exceeds 70%. The narrow gap between the best and worst facility (only 4.8 percentage points) confirms that this is a programme-wide challenge, not a problem with one or two facilities requiring a programme-level governance response.

Discussion

4.1 The core problem: when missing data becomes an AI predictor

The central finding of this study is both simple and consequential: the AI model built on this dataset learns to predict HIV outcomes partly by detecting who has missing demographic records, not solely by detecting who is clinically at risk. This happens because the fields with the worst completeness — education level (1.2–7.9%) and household income (4.7–19.8%) are precisely the fields whose missing-data indicators become the strongest SHAP predictors in the model [6]. When the model is asked 'which features most reliably predict unsuppressed viral load?', the truthful answer, given the data, is: 'whether the patient's education level is recorded'.

This matters for public health equity. If a clinical decision support tool generates an alert for a patient primarily because their education record is blank, the clinician receiving that alert cannot determine whether to respond with a health literacy intervention (the right response if the patient genuinely lacks formal education) or with a data quality correction (the right response if the field is simply blank by

default). Using the wrong response misallocates scarce clinical resources. At scale, across thousands of patients, this systematic misallocation represents a meaningful equity concern, directly relevant to SDG 10 (Reduced Inequalities).

This is not a problem unique to this study or to Zambia. Any AI model trained on EHR data with systematic demographic missingness in resource-limited settings will face the same challenge. The missing-data mechanism is non-random, certain facilities and certain time periods have lower completeness for certain fields and non-random missingness creates exploitable patterns that ML models will detect. Our missingness-outcome confounding test (Table 6, dimension D8) is a general tool for identifying this problem before AI deployment.

4.2 Coverage collapse: why engagement features fail in this context

The 12.2% engagement feature coverage (file2) and 3.3% laboratory feature coverage (file4) reveal a structural limitation of using longitudinal EHR data for AI in programmes with a long paper-based history. The SmartCare system has been in operation since 2004, but real-time electronic data entry only became standard around 2015 [4]. The result is that 87.8% of patients in the AI training dataset have no pre-enrolment digital visit history, their engagement patterns, appointment attendance, and clinical trajectory are invisible to the model.

This severely limits what the AI can learn about behaviour-related risk factors. Treatment interruption and viral suppression are both driven substantially by engagement patterns whether patients attend appointments, whether they collect medications regularly, how their clinical indicators trend over time. An AI model that has no access to engagement history for 87.8% of its training data is working with one arm tied behind its back on the most clinically relevant risk dimensions. The finding that demographic missing-data flags dominate the model's predictions may partly reflect this: in the absence of engagement signals, the model gravitates towards whatever stable signals are available, including missing-data patterns.

4.3 Temporal patterns and sustainable data quality

The temporal completeness analysis (Fig. 3) reveals a warning central to sustainable AI deployment: field-level data quality does not improve passively as systems mature. The data quality problem cannot be solved by system upgrades alone, it requires changes to the clinical workflow and data entry culture at the point of care.

The implication for AI is direct: any AI model trained on this data must be regularly retrained and monitored, and each retraining cycle should include a fresh DQA check. If field completeness for key features changes between training cycles, either improving (after a data quality intervention) or declining (after a campaign ends), the model's predictions for those features may change in ways that are invisible in standard performance metrics but visible in the DQA outputs.

4.4 Eight requirements for trustworthy AI in resource-limited public health

Table 7 presents eight minimum data quality requirements grounded in the empirical findings of this study. Together, they constitute a practical checklist that any AI team working with routine EHR data in similar contexts should complete before training a model.

Table 7

Eight minimum data quality requirements for trustworthy AI trained on routine HIV EHR data in resource-limited settings

ID	Requirement	What to do	Why (Evidence from this study)
R1	Feature coverage threshold	Verify that at least 30% of training patients have usable values for each feature group before including it in the model. Groups below this threshold must be excluded and the exclusion documented.	Applied to file2 (12.2% coverage) and file4 (3.3% coverage) — both are below this threshold in the current dataset.
R2	Appointment lateness plausibility gate	Cap max_days_late and mean_days_late at ± 730 days before model training. Report the number of records affected.	54.3% of patients with engagement data exceeded this limit; the maximum value recorded was 45,459 days (124 years).
R3	Outcome label audit	Compute outcome prevalence per facility before and after restricting records to the relevant time window. Document the source field used for each outcome label.	Enables cross-study comparability and reveals facility-level coding inconsistencies before they enter model training.
R4	Missingness-outcome confounding test	For each missing-data indicator that appears in the model's top-10 SHAP features, compute the Pearson correlation between the facility-level missingness rate and the facility-level outcome rate. Flag $\|r\| > 0.5$ for review.	Education vs UVL ($r = -0.865$) and HIV stage vs TB ($r = +0.741$) exceed this threshold in this study.
R5	Temporal completeness audit	Stratify field completeness by patient enrolment year. If a key field is below 20% completeness for patients enrolled before a specific year, exclude those records from training for that feature.	Education completeness was below 1% for patients enrolled 2004–2015, rising gradually after 2022.
R6	Schema documentation	Document all cases where a data field is present in some facilities but absent in others. Replace generic code workarounds ('if column exists') with facility-specific imputation rules.	23 such workarounds were required in the current pipeline.
R7	Mandatory demographic entry	Configure the EHR system to require a response for education, household income, and marital status, with an explicit 'Not Available' option that is different from a blank or missing entry.	Prevents continued accumulation of the missing-data confounding problem at source.
R8	Standardised outcome definitions	Define IIT12m using the PEPFAR TX_ML lost-to-follow-up indicator. Document the laboratory test name matching rules for UVL12m. Publish both before data collection.	Enables meaningful comparison of AI performance statistics across studies in sub-Saharan Africa.

Three of these requirements, R4 (missingness-outcome confounding test), R1 (feature coverage threshold), and R2 (plausibility gate) directly address the three most serious data quality problems identified in this study. The remaining five address structural issues that are relevant to any programme considering AI deployment: label standardisation (R3, R8), temporal quality auditing (R5), schema documentation (R6), and point-of-care data governance (R7).

Requirements R7 and R8 in particular speak to a broader principle: AI reliability in public health is ultimately a data governance problem, not just a machine learning problem. The most impactful interventions for improving the trustworthiness of AI in this setting are not algorithmic, they are changes to what data is collected, how it is recorded, and how consistently definitions are applied across facilities and over time.

4.5 Practical recommendations for SmartCare

Four changes to SmartCare data governance could be implemented without any new infrastructure investment and would substantially improve the data quality conditions for AI. First, making education, household income, and marital status required fields at patient registration with an explicit 'Not Available' option distinguishable from a blank, would prevent continued accumulation of missing-data confounding at source. Second, adding a validation rule rejecting appointment dates more than one year from the corresponding visit date would eliminate the plausibility violations that affect 54.3% of patients with engagement data. Third, an automated monthly data quality dashboard in the national reporting system would allow programme managers to track completeness rates for AI-critical fields at each facility and target remediation where it is most needed. Fourth, adopting a standardised treatment interruption coding protocol aligned with the PEPFAR TX_ML indicator would resolve the outcome label instability documented in Section 3.3.

4.6 Limitations

This study covers six facilities in Lusaka District; the data quality profile in rural, private, or district-level facilities may differ. We have not conducted interviews with data entry staff to understand the root causes of specific completeness patterns. The missingness-outcome confounding test (D8) uses facility-level correlations across only six data points, which limits statistical power and cannot detect within-facility confounding effects. The DQI is an unweighted composite; a weighted version prioritising features by their AI importance could produce different facility rankings and would be worth exploring in future work.

Conclusions

Routine HIV EHR data in Lusaka, Zambia exhibits three interacting data quality problems that substantially undermine the trustworthiness and fairness of AI models trained on it. Sociodemographic fields that are most important for AI — education level and household income are effectively absent as structured data, causing the model to use their missing-data indicators as predictors instead. When these missing-data patterns are correlated with facility-level outcome rates, the model's predictions

partly reflect administrative rather than clinical differences between patients. Engagement and laboratory features collapse to near-zero coverage, forcing the model to rely almost entirely on a one-time demographic snapshot for the majority of patients. And implausible values in appointment lateness features — including values of up to 124 years enter the model unchecked, distorting predictions for a feature group that the model considers clinically important.

None of these are insurmountable problems. The eight minimum requirements, the DQA toolkit, and the four SmartCare governance recommendations provide a concrete, actionable pathway from the current situation to one in which AI in Zambia's HIV programme is more reliable, more equitable, and better aligned with the goals of SDG 9 (Innovation and Infrastructure) and SDG 10 (Reduced Inequalities). Making AI trustworthy in public health is not primarily a machine learning challenge, it is a data governance challenge. This study provides the evidence base and the practical tools to begin addressing it.

Abbreviations

AI: Artificial intelligence

ART : Antiretroviral therapy

AUC-PR : Area under the precision-recall curve

AUC-ROC : Area under the receiver operating characteristic curve

CDSS : Clinical decision support system

CIDRZ : Centre for Infectious Disease Research in Zambia

DQA : Data quality assessment

DQI – Data quality index

EHR – Electronic health record

HIV – Human immunodeficiency virus

IIT – Interruption in treatment

IIT12m – 12-month IIT outcome

IQR – Interquartile range

LMIC – Low- and middle-income country

ML – Machine learning

MoH – Ministry of Health, Zambia

PEPFAR – President's Emergency Plan for AIDS Relief

PLHIV – People living with HIV

SDG – Sustainable Development Goal

SHAP – Shapley Additive Explanations

SmartCare – Zambia's national HIV electronic health record system

TB – Tuberculosis

TB12m – 12-month incident TB outcome

TX_ML – PEPFAR lost-to-follow-up indicator

UHC – Urban health centre

UVL – Unsuppressed viral load

UVL12m – 12-month UVL outcome

VL – Viral load

WHO – World Health Organization

Declarations

Ethics approval

Ethical approval for this study was granted by the ZCAS University Ethics Review Committee (Reference No. 2025/11/001). All methods were carried out in accordance with relevant guidelines and regulations, including the Declaration of Helsinki. Authorization to access routine electronic health record data was granted by the Ministry of Health, Zambia.

Consent to Participate

This study used fully de-identified secondary programme data from the SmartCare electronic health record system. Informed consent from individual participants or their legal guardians was not required; a formal waiver of informed consent was granted by the ZCAS University Ethics Review Committee (Reference No. 2025/11/001) due to the retrospective nature of the study, in accordance with applicable national guidelines and regulations governing the use of de-identified routine health data for public health research.

Consent to Publish

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research received no funding.

Author Contribution

JP conceptualised and designed the study, developed the analytical framework and software pipeline, performed all data curation, formal analysis, and modelling, and wrote the original draft of the manuscript. AZ contributed to methodological review, provided supervision throughout the study, and reviewed and edited the manuscript. CN contributed to data interpretation, model validation, and reviewed and edited the manuscript. AK and IC facilitated access to and governance of the routine EHR data from the Ministry of Health, Zambia, coordinated institutional authorisation for data use, and reviewed and edited the manuscript. MC contributed to programme-level data coordination and oversight of EHR data sources and reviewed and edited the manuscript. TS and JM facilitated data access through CIDRZ, contributed to clinical and programmatic interpretation of findings, and reviewed and edited the manuscript. ML provided institutional data oversight, contributed to research governance and supervision, and reviewed and edited the manuscript. All authors read and approved the final version of the manuscript.

Acknowledgement

The authors acknowledge the Ministry of Health, Zambia, CIDRZ, and the clinical and data entry staff at all six study facilities. Special acknowledgement to the 246,053 people living with HIV whose care records underpin this study. We also acknowledge the SmartCare programme management team for their commitment to data-driven health system strengthening.

Data Availability

The datasets used during the study are not publicly available due to patient confidentiality requirements and Ministry of Health, Zambia data governance policies. The datasets may be made available from the

corresponding author on reasonable request and with explicit permission from the Ministry of Health, Zambia.

References

1. UNAIDS, Global HIV. & AIDS statistics — fact sheet. 2024. https://www.unaids.org/en/resources/fact-sheet
2. Ngcobo S, et al. Artificial intelligence for HIV care: a global systematic review of current studies and emerging trends. J Int AIDS Soc. 2025;28(10):e70045. https://doi.org/10.1002/jia2.70045.
3. Moomba K, et al. Effects of real-time electronic data entry on HIV programme data quality in Lusaka, Zambia. Public Health Action. 2020;10(1):47–52.
4. Gumede-Moyo S, et al. A qualitative inquiry into implementing an electronic health record system (SmartCare) for prevention of mother-to-child transmission data in Zambia: a retrospective study. BMJ Open. 2019;9(9):e030428.
5. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–53. https://doi.org/10.1126/science.aax2342.
6. Phiri J, Zimba A, Njovu C et al. An explainable multi-outcome machine learning framework for personalized clinical decision support in HIV care using routine electronic health records in Zambia. BMC Med Inf Decis Mak. 2026 [in submission].
7. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment. J Am Med Inf Assoc. 2013;20(1):144–51. https://doi.org/10.1136/amiajnl-2011-000681.
8. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
9. Salami D, Koech E, Turan JM, et al. Prediction of first and multiple antiretroviral therapy interruptions in people living with HIV. JMIR Med Inf. 2026;14:e78964. https://doi.org/10.2196/78964.
10. Ijaiya M, Troncoso E, Mutloatse M, et al. Use of machine learning in predicting continuity of HIV treatment in selected Nigerian states. PLOS Glob Public Health. 2025;5(4):e0004497. https://doi.org/10.1371/journal.pgph.0004497.
11. Abera NM, Alemu TG, Agegnehu CD. Incidence and predictors of virological failure among HIV-infected children on first-line ART. PLoS ONE. 2023;18(11):e0289095. https://doi.org/10.1371/journal.pone.0289095.
12. Mamo DN, 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 Inf Decis Mak. 2023;23:75.
13. Kapoor S, Narayanan A. Leakage and the reproducibility crisis in machine-learning-based science. Patterns. 2023;4(9):100804. https://doi.org/10.1016/j.patter.2023.100804.

14. Peltzer K, Pengpid S. Antiretroviral treatment adherence among HIV patients in rural South Africa. J Int AIDS Soc. 2010.
15. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58. https://doi.org/10.1056/NEJMra1814259.
16. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. Nat Med. 2019;25:1337–40. https://doi.org/10.1038/s41591-019-0548-6.
17. SMART Zambia Institute. SmartCare Pro — advancing healthcare. 2024. https://www.szi.gov.zm
18. Ministry of Health, Zambia. Zambia consolidated guidelines for treatment and prevention of HIV infection. Lusaka: MoH; 2022.
19. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, Collins GS. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016;353:i3140.
20. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. ML4H. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. Proceedings of Machine Learning Research (2019). 2019;106:1–21.

Figures

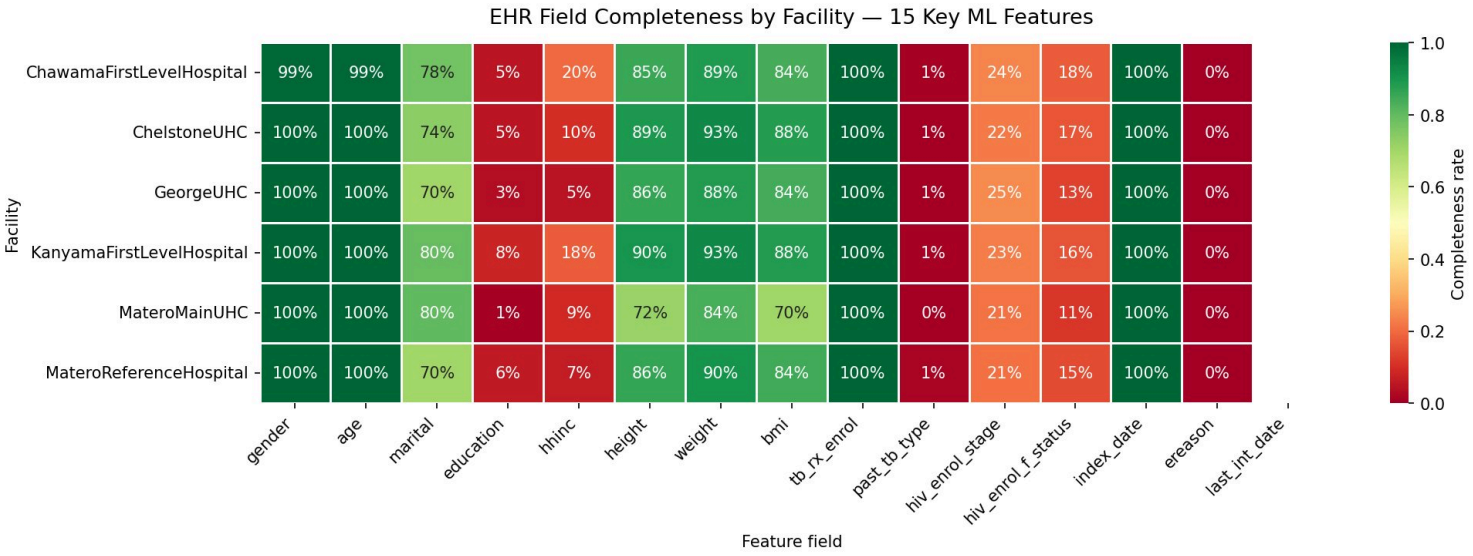


Figure 1

EHR field completeness heatmap.

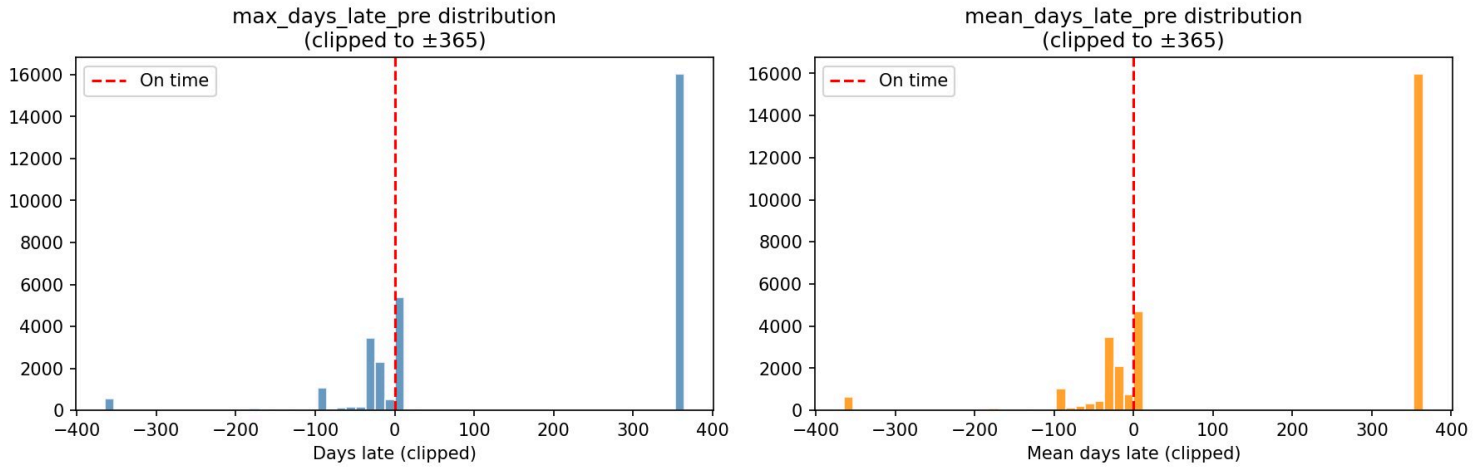


Figure 2

Distribution of appointment lateness features

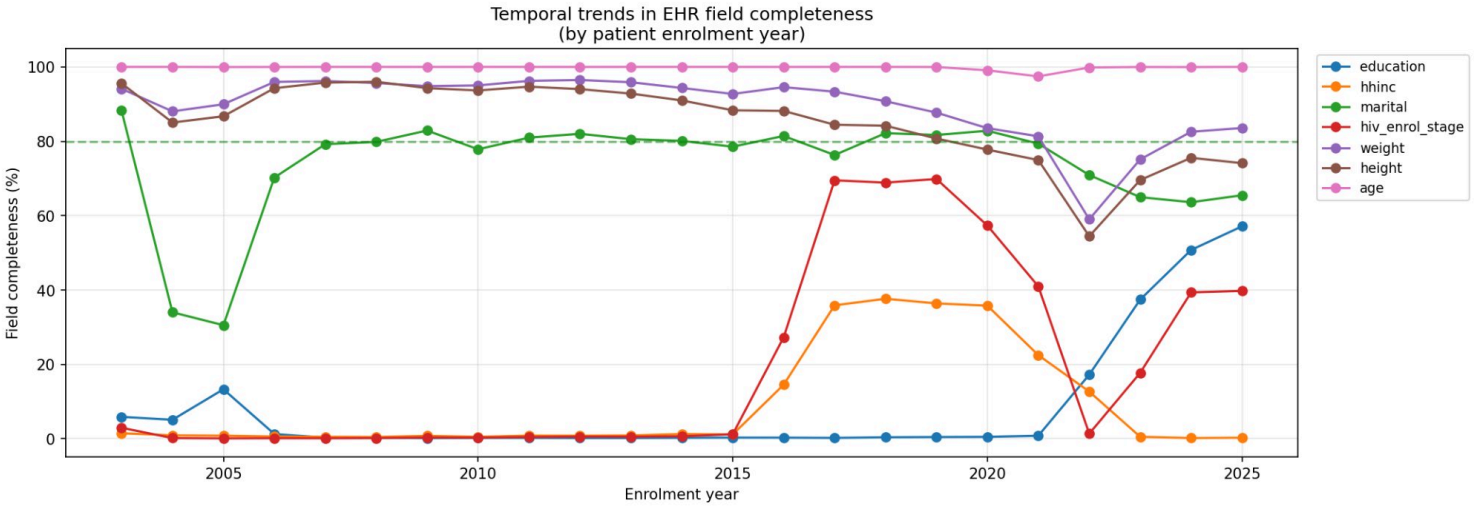


Figure 3

Temporal trends in EHR field completeness by patient enrolment year (2003 – 2025).

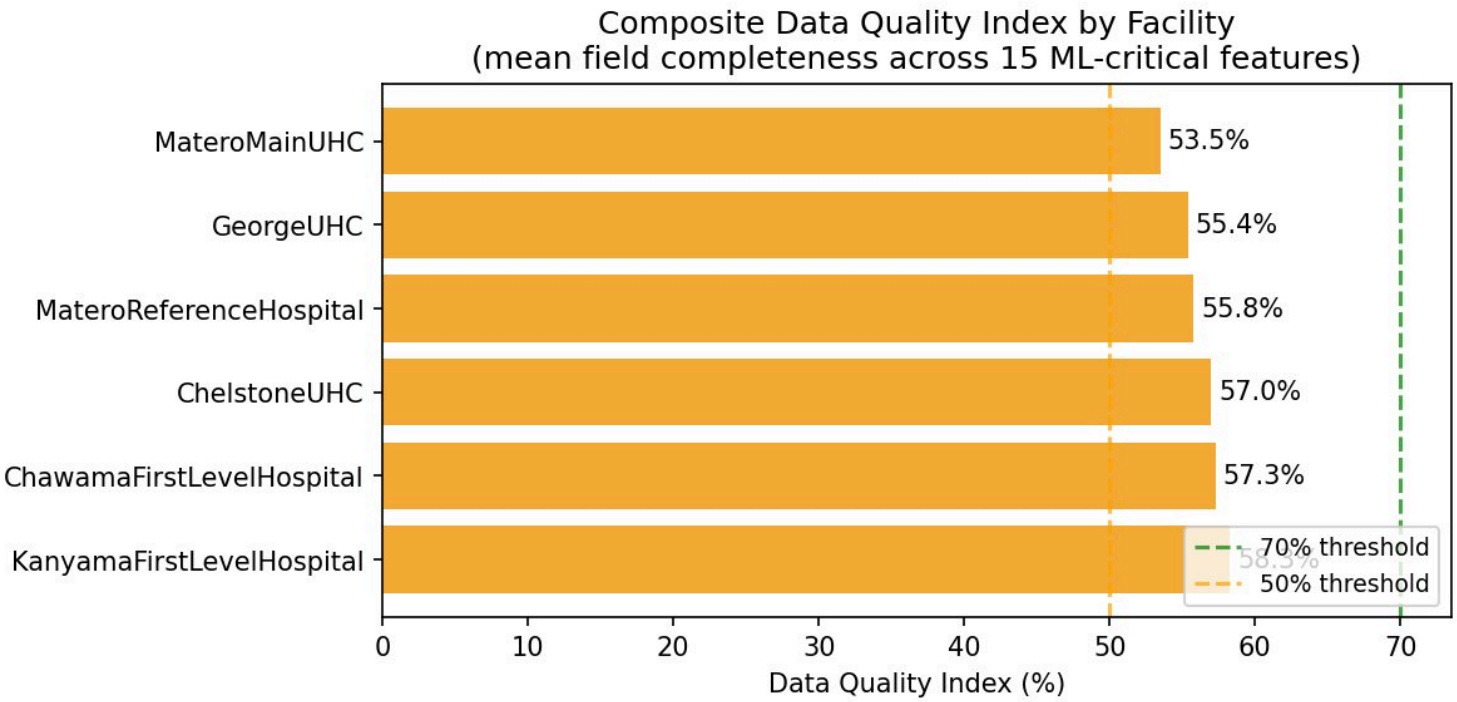


Figure 4

Composite Data Quality Index (DQI) by facility

Supplementary Files

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

- SupplementaryMaterial.docx