ZCAS UNIVERSITY

Lusaka, Zambia

OPTIMIZING MACHINE LEARNING MODELS FOR DISEASE PROGRESSION PREDICTION IN TB/HIV COINFECTION USING RESOURCE-LIMITED EHR DATA SYSTEMS IN ZAMBIA

By

JOE PHIRI

202305036

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy in Computing and Data Science

School of Computing, Technology and Applied Sciences

MAY 2026

Declaration

Name: Joe Phiri

Student Number: 202305036

I hereby declare that this thesis is the result of my own original work, except for quotations and summaries which have been duly acknowledged through citation. This work has not been submitted, in whole or in part, to any other university or institution for any qualification. I give consent for this thesis to be made available for use in the ZCAS University Library and for scholarly purposes in accordance with University policy.

Plagiarism check: _____ %

AI Similarity: _____ %

Signature: _____ Date: _____

Supervisors' Certification

We certify that this thesis was carried out under our supervision and is ready for submission and examination.

Principal Supervisor Name: Professor Aaron Zimba, PhD

Supervisor Signature: _____ Date: _____

Co-Supervisor Name: Dr Chiyaba Njovu

Supervisor Signature: _____ Date: _____

Responsible AI-Use Disclosure

In the interest of transparency and in accordance with the World Association of Medical Editors (WAME) recommendations on the responsible use of artificial intelligence, the following disclosure is provided. Generative AI tools were used in a limited, assistive capacity during the preparation of this thesis, specifically for language editing, improving readability, formatting assistance and consistency checking of prose. All research design, data engineering, modelling, analysis, interpretation and conclusions are the candidate's own intellectual work.

No AI system is credited as an author, and no AI system generated the empirical results, statistical analyses or scientific claims presented herein. The candidate takes full responsibility for the accuracy, integrity, attribution and originality of the entire thesis, including any content that was language-edited with AI assistance. All sources are cited in accordance with IEEE referencing conventions, and the work has been checked for plagiarism and AI similarity as recorded on the Declaration page.

Abstract

The integration of artificial intelligence (AI) and machine learning (ML) into electronic health record (EHR) systems offers a transformative Fourth Industrial Revolution (4IR) opportunity for public health decision-making in resource-limited settings. The translation of algorithmic advances into equitable, reliable and actionable clinical tools in low- and middle-income countries (LMICs), however, remains constrained by data quality deficiencies, methodological gaps and governance challenges. This thesis addresses these challenges through three inter-related empirical studies conducted in the context of Zambia's national HIV EHR system (SmartCare), drawing on a deduplicated, deterministically linked dataset of 246,053 people living with HIV (PLHIV) across six high-volume public health facilities in Lusaka District.

Study One presents a systematic review of 64 peer-reviewed studies published between 2018 and 2025 on ML-based predictive analytics using EHR data, with LMICs as the primary focus. Conducted in accordance with PRISMA 2020 guidelines and using a PROBAST-AI-adapted risk-of-bias instrument, the review establishes that only 18.8% of studies were conducted in LMIC settings despite these settings carrying the greatest burden of preventable disease. External validation was reported in only 7.8% of studies, calibration in 6.2%, and explainability in 8.3% of LMIC-focused studies compared with 36.4% in high-income settings. Four cross-cutting evaluation, contextual, ethical and deployment gaps are identified and used to scope the empirical chapters.

Study Two develops and evaluates a leakage-safe, multi-outcome, externally validated and explainable ML framework predicting three 12-month binary outcomes among PLHIV: recorded TB treatment (TB12m), programme-recorded interruption in treatment (IIT12m) and recorded unsuppressed viral load (UVL12m). A stacked ensemble of logistic regression, random forest and XGBoost with a logistic regression meta-learner achieved AUC-ROC of 0.707 for TB12m, 0.839 for IIT12m and 0.778 for UVL12m on a held-out facility test set (n = 74,506). F1-optimised decision thresholds support operational protocols ranging from high-sensitivity screening (TB12m, threshold 0.257, sensitivity 92.3%) to highly targeted intensive intervention (IIT12m and UVL12m, thresholds near 0.9). SHAP-based local interpretability shows that education level is the dominant predictor of UVL12m and a leading predictor of IIT12m, raising critical questions about whether the signal reflects genuine clinical risk or facility-level data recording patterns. Facility-holdout validation confirms cross-site generalisability under deployment-realistic conditions.

Study Three conducts a systematic eight-dimension data quality assessment (DQA) of the SmartCare EHR system using the AI pipeline as a diagnostic lens. Education-level completeness ranged from 1.2% to 7.9% across all six facilities; engagement feature coverage was only 12.3%; and 54.3% of patients with engagement data had implausible appointment-lateness values, the highest being 45,459 days (approximately 124 years). The composite Data Quality Index (DQI) ranged from 53.5% to 58.3%, all below the 70% minimum threshold proposed in the thesis for

fully trustworthy AI training. Eight minimum data quality requirements and a reproducible DQA toolkit are proposed, with quantitative remediation targets at facility and programme levels.

Complementing these empirical studies, the thesis develops a 4IR health transformation framework situating AI adoption within Zambia's broader digital health ecosystem; constructs an economic value creation model demonstrating positive cumulative cost-benefit by Year 3 of deployment under base-case assumptions; and proposes an institutional governance architecture together with a five-phase AI adoption roadmap (2025-2030) aligned with PEPFAR, Global Fund and Ministry of Health programmatic cycles. A proposed ICT integration architecture linking SmartCare to an HL7 FHIR R4-compliant ML prediction engine, DHIS2, DISA and a clinical decision support dashboard provides the engineering blueprint for transitioning from research-grade pipelines to production-ready clinical deployment at national scale.

Two formal dissemination meetings were convened during the research: with the Ministry of Health Provincial Health Office Monitoring and Evaluation Data Management teams (representatives from Lusaka, Southern, Western, North-Western and Eastern Provinces); with the AI Think-Tank Team of experts at the Centre for Infectious Disease Research in Zambia (CIDRZ). Feedback across these forums was uniformly positive and is integrated into the thesis recommendations.

Together, these contributions show that trustworthy AI in resource-limited public health requires as much investment in data governance, economic justification, ICT architecture and institutional capacity as in algorithmic development. The thesis offers an integrated framework, evidence base and operational roadmap for responsible AI deployment in Zambia's HIV programme, with direct relevance to comparable settings across sub-Saharan Africa.

Keywords: *Machine learning; electronic health records; predictive analytics; HIV; tuberculosis; data quality; explainable AI; Fourth Industrial Revolution; low- and middle-income countries; SHAP; stacked ensemble; algorithmic fairness; clinical decision support; ICT architecture; FHIR; governance; equity; Zambia; sub-Saharan Africa.*

Acknowledgements

This thesis would not have been possible without the professional and institutional support of many individuals and organisations, to whom I wish to express my sincere gratitude.

I am deeply indebted to my supervisors, Professor Aaron Zimba and Dr Chiyaba Njovu, for their intellectual mentorship, methodological rigour and sustained encouragement throughout this doctoral journey. Their guidance has shaped both the research and my growth as a scholar, setting a standard for principled, evidence-grounded supervision.

I gratefully acknowledge the Ministry of Health, Zambia, particularly Mr Andrew Kashoka and Mr Innocent Chiboma, for granting institutional authorisation to access the SmartCare programme data and for their commitment to evidence-based health systems strengthening.

I thank the Centre for Infectious Disease Research in Zambia (CIDRZ), especially Mr Trevor Sinkala, Mr Jacob Mutale and Mr Mwansa Lumpa, for facilitating data access, providing epidemiological expertise and supporting research governance. The CIDRZ AI Think-Tank Team provided invaluable cross-disciplinary feedback that shaped the equity safeguards and governance architecture proposed in this thesis.

I am grateful to Mr Mulenga Chiwele and the SmartCare programme data and monitoring and evaluation teams at district and provincial levels, and to the provincial monitoring and evaluation representatives from Lusaka, Southern, Western, North-Western and Eastern Provinces who participated in the first dissemination meeting, whose questions and reflections substantially sharpened the recommendations and roadmap.

I acknowledge the clinical and data staff at Chawama First Level Hospital, Chelstone Urban Health Centre, George Urban Health Centre, Kanyama First Level Hospital, Matero Main Urban Health Centre and Matero Reference Hospital, whose diligence in patient-level documentation made this work possible.

Finally, I acknowledge the 246,053 people living with HIV whose de-identified clinical records form the empirical foundation of this thesis; this work is undertaken in the hope of improving care for every one of them.

This research received no external funding. All computational analyses were conducted on personal and university-supported computing resources.

Dedication

To my wife and children, for the time given up, the holidays postponed and the quiet evenings during which this work took shape; your patience, encouragement and love sustained this entire journey.

To my parents and siblings, whose faith in this journey predated the journey itself, and whose steadfast moral support carried me through its most demanding moments.

To my friends and all who offered encouragement and quiet support along the way, and in gratitude for the faith that has been my constant source of strength and perseverance.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Abbreviation	Full Meaning
4IR	Fourth Industrial Revolution
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
FHIR	Fast Healthcare Interoperability Resources
HIV	Human Immunodeficiency Virus
IIT12m	Interruption in Treatment within 12 Months
LLM	Large Language Model
LMIC	Low- and Middle-Income Country
ML	Machine Learning
PLHIV	People Living with HIV
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SHAP	SHapley Additive exPlanations
TB12m	Incident Tuberculosis within 12 Months
UVL12m	Unsuppressed Viral Load within 12 Months
WHO	World Health Organization
XGBoost	Extreme Gradient Boosting

CHAPTER 1: INTRODUCTION

1.1 Background to the Study

1.1.1 Global Burden of Disease and the LMIC Health Information Gap

The global burden of preventable and manageable disease continues to fall disproportionately on low- and middle-income countries (LMICs), where health systems operate under severe and persistent resource constraints. Sub-Saharan Africa, in particular, bears the heaviest burden of HIV/AIDS, tuberculosis (TB) and a growing epidemic of non-communicable diseases (NCDs), while simultaneously contending with fragmented health information infrastructure, limited analytical capacity and chronically underfunded health services [1], [2]. The World Health Organization estimates that more than 25 million people are living with HIV in sub-Saharan Africa, and that the region accounts for approximately 70% of new infections each year [2]. TB co-infection compounds the clinical and operational complexity of HIV care, with PLHIV facing a 20-to-30-fold higher risk of developing active TB than the general population [3].

These epidemiological realities meet a system-level reality that is equally consequential: the chronic underdevelopment of health information infrastructure. In Zambia, only 883 of 3,540 public health facilities (25%) have reliable internet connectivity, and only a subset of those operate a fully digital electronic health record (EHR) system [28]. Among facilities that do collect digital records, data are typically structured around programme reporting requirements rather than research-grade longitudinal documentation, with the result that data quality is heterogeneous, schema definitions drift across facilities and time, and large fractions of the patient population have incomplete demographic and clinical records [3], [4], [13]. The combination of high disease burden and weak information infrastructure creates a particularly acute version of a global problem: the populations with the greatest potential to benefit from data-driven care improvements are precisely the populations whose data are least well captured.

This thesis is positioned at the intersection of these two realities. It asks how the application of artificial intelligence (AI) and machine learning (ML) to routine HIV EHR data in Zambia can be designed, evaluated and deployed in a way that produces operationally meaningful,

methodologically defensible and ethically responsible decision support, without overstating what the data can support and without exacerbating the equity problems that the data themselves embed.

1.1.2 The Fourth Industrial Revolution and its Health Sector Promise

The Fourth Industrial Revolution (4IR), characterised by the convergence of artificial intelligence, machine learning, big data analytics, the Internet of Things and cloud computing, is reshaping industries and public services worldwide [1], [2]. In healthcare, 4IR technologies offer the prospect of precision medicine, predictive surveillance and intelligent resource allocation at a scale previously unachievable [3], [4]. Schwab's conceptualisation of the 4IR identifies three converging technology clusters: physical systems such as autonomous vehicles and advanced robotics; digital systems such as AI, big data and the IoT; and biological systems such as genomics and synthetic biology [1]. In health systems, the most immediately consequential cluster for LMICs is the digital one, and particularly AI and ML applied to routine health data [3], [4].

For sub-Saharan Africa, the 4IR represents both an opportunity to leapfrog traditional technological development pathways and a risk of entrenching inequalities if adoption lacks contextual grounding, institutional readiness and equitable governance [5], [6]. Examples of successful African 4IR health adoption include Rwanda's use of drone technology for blood-product delivery; South Africa's deployment of ML for tuberculosis diagnosis from chest X-rays; and Kenya's integration of mobile health platforms into community health systems [17], [18]. These cases share a recognisable pattern: successful adoption is preceded by substantial investment in foundational infrastructure (connectivity, digital literacy, system integration) and governance (regulatory frameworks, data protection laws, ethics oversight) [5], [17].

Zambia's HIV programme represents one of the most mature digital health contexts in sub-Saharan Africa in which 4IR analytics can be applied responsibly. The country's SmartCare HIV EHR has been in operation since 2004 and is deployed across more than 1,600 public health facilities nationwide, enrolling over two million patients [3], [4]. SmartCare anchors a broader ecosystem that includes the District Health Information System 2 (DHIS2) for aggregate reporting, the District Information System (DISA) for laboratory information, the Electronic Logistics Management Information System (ELMIS) for pharmaceutical supply chain data and the Data for Accountability, Transparency and Impact Monitoring (DATIM) system for President's Emergency Plan for AIDS Relief (PEPFAR) reporting. Together these systems generate millions of patient-

level longitudinal records annually, providing the raw material on which AI-enabled decision support can be built.

1.1.3 The Zambia HIV Context and Patient Population

Zambia has achieved notable programmatic progress against HIV/AIDS. As of 2024, 98% of PLHIV are aware of their status, 98% of those diagnosed are receiving antiretroviral therapy (ART), and 97% of those on ART are virally suppressed, exceeding the UNAIDS 95-95-95 targets [29]. These achievements, made possible by sustained investment from the Ministry of Health, PEPFAR, the Global Fund and other partners, have transformed HIV from an acute fatal illness into a chronic, manageable condition for the majority of patients in care. The remaining HIV programme priorities, however, are precisely those for which AI-enabled prediction can add the most value: identifying which patients are at risk of interrupting treatment, which are at risk of unsuppressed viral load and which are at risk of developing TB co-infection, before adverse outcomes occur.

The analytical dataset on which this thesis is grounded comprises 246,053 deduplicated PLHIV ever recorded in SmartCare across six high-volume Lusaka District facilities: Chawama First Level Hospital, Chelstone Urban Health Centre, George Urban Health Centre, Kanyama First Level Hospital, Matero Main Urban Health Centre and Matero Reference Hospital. The cohort has a median age of 32 years (IQR 26–38) and is 62.5% female, broadly consistent with the demographic profile of Zambia's national HIV programme [29]. Enrolments span 2003 to 2025, giving more than two decades of longitudinal data depth in the four facilities with the longest digital history.

1.1.4 Why Methodological Rigour Matters

Despite the apparent volume of EHR data available, translation of AI and ML advances from research to routine public health practice has been uneven. The methodological literature is dominated by studies conducted in high-income settings with large, well-curated EHR datasets, advanced computational infrastructure and mature data governance frameworks [9], [10]. Methods developed and validated in such contexts do not transfer directly to LMIC health systems, where data quality is systematically compromised, digital infrastructure is unreliable and the

interpretability and equity implications of algorithmic outputs carry heightened consequences for underserved populations [11], [12].

Several specific methodological problems plague the published literature on ML applied to LMIC EHR data, each of which is examined empirically in subsequent chapters. First, external validation is reported in only a small minority of studies, meaning that performance claims often do not generalise beyond the dataset on which the model was fit. Second, calibration assessment, which evaluates whether predicted probabilities correspond to real-world event rates, is systematically underreported despite being essential for risk-stratification and resource-allocation decisions. Third, explainability assessment is reported in only 8.3% of LMIC-focused studies, far below the 36.4% rate in high-income studies, raising concerns about whether models can be safely deployed in workflows that require clinician trust. Fourth, the temporal feature engineering that prevents information leakage is poorly documented in many published studies, leading to inflated apparent performance that collapses on deployment.

A further dimension, often neglected in the ML literature but central to this thesis, is the quality of the input data itself. EHR data quality in resource-limited settings is rarely systematically assessed before model development, and the field lacks an agreed framework for evaluating whether routine data are fit for AI training. The work presented in this thesis demonstrates that, when the data quality lens is applied rigorously, several conventional assumptions about AI performance must be revised, with consequences for how models should be interpreted, deployed and governed.

1.1.5 The Equity Dimension

AI in public health is not a neutral technical project. Algorithms trained on biased or incomplete data can systematically disadvantage marginalised populations whose records are sparser or whose clinical patterns differ from those represented in the training data [47], [53]. Obermeyer and colleagues showed that a commercial algorithm used to allocate health resources in the United States systematically underestimated the needs of Black patients, because the proxy used to define need (historical health spending) reflected structural inequalities in access to care rather than underlying illness [47]. Equivalent risks apply to AI in HIV care in Zambia: the dominant feature in predicting unsuppressed viral load, as Chapter 5 demonstrates, is education level, and the completeness of this field varies sharply across facilities and population strata, with the lowest completeness at the facilities serving the most deprived populations.

This thesis takes seriously the equity dimension of AI adoption. The governance architecture proposed in Chapter 6 includes algorithmic fairness audits as a mandatory model-approval gate; the deployment workflow includes human-in-the-loop review for cases flagged primarily on the basis of missing data; and the proposed AI Governance Committee includes explicit civil-society and patient representation. The argument throughout is that responsible AI adoption in resource-limited public health requires the equity dimension to be designed into the architecture, not added on after deployment.

1.1.6 Why this Thesis, Why Now

Three convergent factors make the present moment the right one for the integrated research programme reported in this thesis. First, the SmartCare EHR has reached sufficient longitudinal depth (now over two decades of patient-level data in the longest-established facilities) and sufficient national coverage (over 1,600 facilities, more than two million patients) to support meaningful ML research at scale [3], [4]. Second, the methodological infrastructure for trustworthy AI — PRISMA 2020, PROBAST-AI, TRIPOD-AI, SHAP-based interpretability frameworks — has matured to the point where it can be deployed within LMIC research workflows without prohibitive technical overhead [20], [21], [22], [52]. Third, the policy environment in Zambia, including the Data Protection Act of 2021, the National Health Strategic Plan and active engagement by the Ministry of Health on digital health transformation, has created the institutional conditions in which evidence-based recommendations from a doctoral thesis can plausibly translate to operational change [80].

1.2 Problem Statement

Despite the growing body of research applying ML to EHR-based health prediction, four interrelated gaps persist in the literature and in practice, with particular acuity in LMIC settings. Together, these gaps create a situation in which AI tools may be deployed in public health settings without adequate evidence of their reliability, fairness, generalisability or institutional sustainability — with potentially serious equity consequences for the populations they are designed to serve.

1.2.1 Gap 1: Methodological Rigour

Methodological rigour in the published literature is systematically deficient. Models are evaluated primarily on discrimination metrics such as the area under the receiver operating characteristic curve (AUC-ROC), which does not assess the probability reliability essential for public health resource allocation decisions, and external validation, calibration and explainability are reported only rarely in LMIC-focused studies [14], [15]. A related dimension, especially consequential for LMIC research, is information leakage: where contemporaneous or post-index information enters the features, apparent performance is inflated and models collapse on prospective deployment.

1.2.2 Gap 2: Data Quality as a Precondition for Trustworthy AI

Data quality as a precondition for trustworthy AI is rarely assessed before model development. The field lacks a systematic framework for judging whether routine EHR data in resource-limited settings is fit for AI training [16]. Existing frameworks were developed for high-income clinical research settings and do not address the dimensions of data quality most consequential for AI, in particular the risk that predictions are driven by data-completeness patterns rather than by clinical signal. How severe this problem is in a national LMIC HIV programme, and how it should be diagnosed, remains an open question.

1.2.3 Gap 3: Multi-Outcome, Externally Validated Frameworks

Multi-outcome, leakage-safe and externally validated frameworks for clinical decision support in HIV care are absent from the LMIC literature. Existing studies typically predict a single outcome in isolation [49], [50], [64], yet HIV clinical decision making necessarily involves multiple competing priorities simultaneously. A patient presenting at a routine visit may be at risk of TB co-infection, treatment interruption and unsuppressed viral load at the same time; an effective decision support system must integrate predictions across these outcomes, not treat them as separate prediction tasks with independent thresholds and workflows.

Existing studies also typically rely on random patient-level train-test splits, which do not test the cross-facility generalisation that deployment in a health system requires.

1.2.4 Gap 4: The ICT Deployment Gap

The ICT architecture required to operationalise ML frameworks in routine clinical settings has not been systematically addressed in the LMIC literature. A critical gap exists between research-grade ML pipelines, which operate as offline analytical scripts on static datasets, and production-ready clinical decision support systems (CDSS) that operate continuously within the ICT infrastructure of a national HIV programme. Bridging this gap requires systematic engineering decisions about data extraction from EHR systems, temporal leakage prevention at the system boundary, scalable preprocessing and model inference, integration with national health IT systems (DHIS2, DISA, ELMIS), clinical dashboard delivery and security, privacy and audit requirements.

This gap is especially acute in Zambia, where limited facility connectivity, power instability and constrained IT support mean that an architecture assuming continuous connectivity would not reach most of the national health network.

1.2.5 Synthesis: Why an Integrated Response is Needed

Taken together, these four gaps create a coherent argument for the integrated, multi-component thesis design adopted here. Methodological rigour alone is insufficient if the data quality foundations are inadequate. Data quality remediation alone is insufficient if the methodology cannot generate trustworthy predictions even from improved data. Trustworthy predictions alone are insufficient if the institutional governance architecture cannot translate them into authorised, audited, equity-protected clinical action. Authorised governance alone is insufficient if the ICT deployment architecture cannot scale predictions to the patients and clinicians who need them. Each component requires the others. The principal contribution of this thesis lies in the integration: showing, through evidence and design, that responsible AI deployment in Zambia's HIV programme is feasible only when all four dimensions are addressed together.

1.3 Aim and Objectives of the Study

The aim of this study is to develop and evaluate an optimised, leakage-safe, externally validated and explainable machine learning framework for predictive analytics in resource-limited electronic health record systems, using Zambia's SmartCare HIV EHR as the empirical context, and to situate

the framework within an integrated 4IR, governance and ICT deployment architecture that supports responsible national-scale adoption.

The aim is operationalised through four specific objectives.

1.3.1 Objective One: Evidence Synthesis

To systematically review and synthesise the international evidence base on ML methods, evaluation practices and deployment challenges for public health prediction using EHR data, with specific emphasis on LMIC settings and methodological rigour beyond discrimination metrics. The review follows the PRISMA 2020 guideline and applies a PROBAST-AI-adapted risk-of-bias instrument to identify the four cross-cutting gaps that scope the empirical chapters.

1.3.2 Objective Two: ML Framework Development and Validation

To develop, validate and evaluate a multi-outcome, leakage-safe and explainable ML framework for predicting clinically actionable HIV care outcomes — recorded TB treatment (TB12m), programme-recorded treatment interruption (IIT12m) and recorded unsuppressed viral load (UVL12m) — from routine EHR data in Zambia, using a facility-holdout external validation design. The framework is to be evaluated on discrimination, threshold-based operational performance and interpretability, with SHAP-based patient-level explanations as a precondition for clinician trust.

1.3.3 Objective Three: Data Quality Assessment

To conduct a systematic, multi-facility data quality assessment of Zambia's national HIV EHR system using an eight-dimension framework, quantify the extent to which data quality failures compromise AI model trustworthiness, and develop minimum data quality requirements and governance recommendations for responsible AI in resource-limited public health. The assessment uses the AI pipeline itself as a diagnostic instrument, surfacing data quality problems that standard statistical checks do not capture.

1.3.4 Objective Four: 4IR Governance and ICT Architecture

To develop a 4IR health transformation framework, an institutional governance architecture and an ICT integration blueprint that collectively provide a pathway from ML research to equitable,

sustainable clinical deployment in Zambia's national HIV programme. The architecture is to include explicit gate criteria for phased adoption, fairness-protective workflow design, and engineering provisions for offline resilience that match the realities of the Zambian health infrastructure. This objective is deliberately integrative rather than narrow: it bundles three interdependent translational deliverables, the governance architecture, the ICT deployment blueprint and the supporting economic case, under a single deployment-pathway aim, because none of the three is meaningful in isolation, a governance regime without an ICT substrate cannot be enacted, and neither can be sustained without an economic rationale. These deliverables are reported separately in the significance and contribution statements (Sections 1.6 and 7.1.1) because each makes a distinct scholarly claim, but they are pursued here as one coherent objective rather than three, so that the four objectives map cleanly onto the four framework components developed in Chapter 5. The asymmetry between four objectives and the larger number of itemised contributions is therefore intentional, and reflects the difference between the unit of research design (the objective) and the unit of scholarly claim (the contribution).

1.4 Research Questions

The research questions are organised around the empirical studies that constitute the thesis, with cross-cutting questions on governance and deployment addressed in the integrative chapters. Each research question is mapped to a specific study and to specific analytical outputs, with the full mapping summarised in Chapter 3.

1.4.1 Study One: Systematic Review

RQ1.1: Which ML algorithms are most frequently applied in LMIC settings for EHR-based prediction, and how do they compare in predictive accuracy, interpretability and computational feasibility?

RQ1.2: Which preprocessing and feature engineering strategies best address data quality challenges in resource-constrained EHR environments?

RQ1.3: How are ML-based predictions evaluated and validated in low-resource settings, particularly regarding calibration and explainability for public health decision-making?

1.4.2 Study Two: ML Modelling Framework

RQ2.1: Can a leakage-safe, multi-outcome ML framework trained on routine HIV EHR data predict TB incidence, treatment interruption and viral load suppression failure with clinically useful discrimination on a held-out facility test set?

RQ2.2: Does a facility-holdout external validation design confirm cross-site generalisability of the ML framework, and how does performance compare with that reported in published LMIC studies using less stringent validation designs?

RQ2.3: What patient-level and population-level feature explanations does SHAP provide, and what are the implications for targeted clinical intervention and equitable resource allocation?

1.4.3 Study Three: Data Quality Assessment

RQ3.1: What are the extent, pattern and mechanisms of data quality deficiencies in Zambia's SmartCare HIV EHR across eight quality dimensions: completeness, conformance, plausibility, feature coverage, outcome sensitivity, temporal stability, cross-facility concordance and missingness-outcome confounding?

RQ3.2: To what extent are AI model predictions confounded by missing-data patterns rather than genuine patient clinical risk, and what is the magnitude of this confounding for the highest-importance features identified by SHAP?

RQ3.3: What minimum data quality requirements and governance reforms are necessary for trustworthy, equitable AI in this setting, and how should they be operationalised at facility, district and national levels?

1.4.4 Cross-Cutting Questions on Governance and Deployment

RQ4.1: What 4IR health transformation framework, institutional governance architecture and economic value-creation model are needed to support sustainable AI deployment in Zambia's HIV programme, and what is the projected cost-benefit trajectory under base-case assumptions?

RQ4.2: What ICT integration architecture and phased national adoption roadmap are required to bridge the gap between research-grade ML pipelines and production-ready clinical decision support at national scale, and how should the architecture be designed for the connectivity and infrastructure realities of Zambian public health facilities?

1.5 Scope and Limitations

Before the individual boundaries of the study are set out, it is important to clarify the intended scope of the work as a whole, because the breadth of the four strands — machine-learning modelling, data-quality assessment, institutional governance and ICT architecture — might be read as attempting too much. These are not four independent studies pursued in parallel; they are four dependent components of a single, integrated pipeline addressing one question: what is required to make predictive analytics trustworthy and deployable in a resource-limited HIV electronic health record system. Data quality is the precondition, the multi-outcome machine-learning framework is the core technical contribution, and the governance and ICT architectures are the translational layer that carries the same models, trained on the same cohort, toward the point of care. The two components examined in the greatest empirical depth are the leakage-safe, externally validated, explainable ML framework (Study Two) and the eight-dimension data-quality assessment (Study One and Study Three); the governance and ICT architectures are advanced as a reasoned, theoretically grounded translational contribution rather than as separately validated empirical studies, and are qualified as such.

The novelty of the thesis lies in this integration rather than in any single component in isolation. Each of the four gaps that motivate the study (Section 1.2) has been noted individually in prior work; what has not previously been done for an LMIC HIV programme is to address them together — combining leakage-safe multi-outcome prediction, facility-holdout external validation, formal data-quality certification as a deployment gate, and an explicit governance and deployment architecture — on a single, large, deterministically linked national cohort. The systematic review (Study One) provides the evidence that these gaps are real and under-addressed in LMIC settings, and thereby anchors the claim to novelty in the literature rather than asserting it.

1.5.1 Geographic Scope

This thesis focuses on HIV-related health outcomes within the context of Zambia's national HIV programme. The empirical studies are limited to six high-volume public health facilities in Lusaka District: Chawama First Level Hospital, Chelstone Urban Health Centre, George Urban Health Centre, Kanyama First Level Hospital, Matero Main Urban Health Centre and Matero Reference Hospital. These six facilities account for a substantial subset of Lusaka District's enrolled HIV

population and provide a deduplicated analytical cohort of 246,053 patients. Generalisability to rural facilities, private facilities and district-level facilities in other provinces requires further investigation, and is identified in Chapter 7 as a priority extension of the work. The Phase 5 of the proposed national adoption roadmap (Chapter 6) sets out the sequenced provincial deployment path that would extend the framework beyond Lusaka District.

1.5.2 Disease Scope

The empirical studies focus on three specific HIV-related outcomes: recorded TB treatment within 12 months of the index date (TB12m), which cannot distinguish incident from prevalent or ongoing treatment; programme-recorded treatment interruption within 12 months (IIT12m); and unsuppressed viral load within 12 months (UVL12m). These were selected because they are clinically actionable, operationally relevant to differentiated service delivery decisions and adequately represented in the routine SmartCare data. The framework is portable to additional outcomes such as advanced HIV disease (AHD), opportunistic infections and longer-horizon mortality, but those extensions are not pursued in this thesis. Generalisability to TB-only contexts, NCDs and maternal-and-child health contexts requires separate investigation.

1.5.3 Methodological Scope

The systematic review is restricted to English-language publications from 2018 to 2025 retrieved from five major bibliographic databases (IEEE Xplore, Scopus, Web of Science, ACM Digital Library and PubMed). Grey literature and non-English-language publications are excluded. The ML framework is not evaluated in a prospective deployment setting; real-time performance may differ owing to data-entry lag, workflow integration challenges and temporal shifts in patient population. The data quality assessment is restricted to the SmartCare EHR system; linked DHIS2, DISA and ELMIS data quality is not separately assessed in this thesis although it is identified as a priority component of any national data governance reform.

1.5.4 Economic and ICT Scope

The economic analysis (Chapter 6) is illustrative rather than facility-level costed; precise cost-benefit calculations would require facility-level costing data and patient-level outcome estimates beyond the scope of this thesis. The CDSS integration architecture is presented as a blueprint for

implementation; it has not yet been prospectively deployed. The Phase 2 of the adoption roadmap (Chapter 6) sets out the shadow-mode pilot deployment at the six Lusaka District study facilities that would constitute the first operational test of the architecture.

1.6 Significance of the Research

This thesis makes several original contributions to health informatics, AI in public health, digital health governance and 4IR policy.

1.6.1 Methodological Contribution

The thesis provides a systematically reviewed and critically appraised synthesis of ML-based EHR predictive analytics with LMIC settings as the primary focus. The review demonstrates that methodological investment in algorithmic novelty has substantially outpaced investment in evaluation rigour, explainability and governance. The PROBAST-AI-adapted risk-of-bias instrument used in the review provides a structured tool that can be applied to subsequent systematic reviews in this area. The thesis also operationalises the facility-holdout external validation design, the leakage-safe temporal feature engineering approach and the eight-dimension DQA framework, all of which are reusable methodological contributions.

1.6.2 Applied Empirical Contribution

The thesis presents a multi-outcome, leakage-safe, facility-holdout validated and SHAP-interpreted ML framework for HIV care outcome prediction from routine Zambian EHR data. With 246,053 patients across six facilities, it is among the largest HIV EHR ML studies in sub-Saharan Africa, and one of the first to combine multi-outcome prediction, leakage-safe temporal feature engineering, facility-holdout external validation and SHAP interpretability in a single integrated pipeline. The framework's AUC-ROC values — 0.707 for TB12m, 0.839 for IIT12m and 0.778 for UVL12m — compare favourably with the published LMIC literature and were obtained under a more stringent validation design than 86% of comparable studies.

1.6.3 Data Governance Contribution

The thesis presents a systematic eight-dimension data quality assessment of Zambia's SmartCare HIV EHR system using an AI pipeline as a diagnostic instrument, producing evidence-based

minimum requirements and a reproducible DQA toolkit applicable across LMIC HIV programmes. The missingness-outcome confounding test (D8) is, to the author's knowledge, the first published operationalisation of this dimension in an LMIC HIV context, and provides the methodological link between data quality findings and ML interpretability findings that subsequent governance recommendations depend on.

1.6.4 4IR Policy and Economic Contribution

The thesis develops a 4IR health transformation framework, an economic value-creation model and an institutional governance architecture that collectively provide the business and policy case for responsible, sustainable AI adoption in Zambia's HIV programme. The five-phase adoption roadmap (2025–2030) aligns AI deployment milestones with Ministry of Health, PEPFAR and Global Fund programmatic cycles. The roadmap is explicitly gated, meaning that progression from each phase to the next is conditional on demonstrated data quality, performance, fairness and operational criteria, rather than on calendar schedule alone.

1.6.5 ICT Engineering Contribution

The thesis proposes a CDSS integration architecture linking SmartCare to an HL7 FHIR R4-compliant ML prediction engine, DHIS2, DISA and a clinical dashboard. This represents a systematic ICT architecture design for AI-enabled clinical decision support in Zambia's national HIV EHR system, with offline-resilient design choices that match the connectivity reality (only 25% of facilities have reliable internet connectivity). The Temporal API Gateway proposed in the architecture operationalises leakage prevention at the system level, not merely at the script level.

1.6.6 Sustainable Development Goal Alignment

The study directly advances Sustainable Development Goal (SDG) 3 (Good Health and Well-being) through improved targeting of clinical interventions in HIV care; SDG 9 (Industry, Innovation and Infrastructure) through advancing digital health infrastructure and AI capability in an African public health setting; and SDG 10 (Reduced Inequalities) through the explicit equity safeguards built into the governance architecture and the fairness-corrected deployment workflow.

1.7 Conceptual Framework

The research is organised around a multi-layer conceptual framework that integrates the four contributions into a coherent response to the challenge of trustworthy AI in resource-limited public health. As illustrated in Figure 1.1, the framework proceeds from a research-context layer (HIV disease burden, SmartCare EHR, 4IR opportunity, resource constraints) through three inter-related empirical study components, and extends to an economic, governance and ICT deployment synthesis.

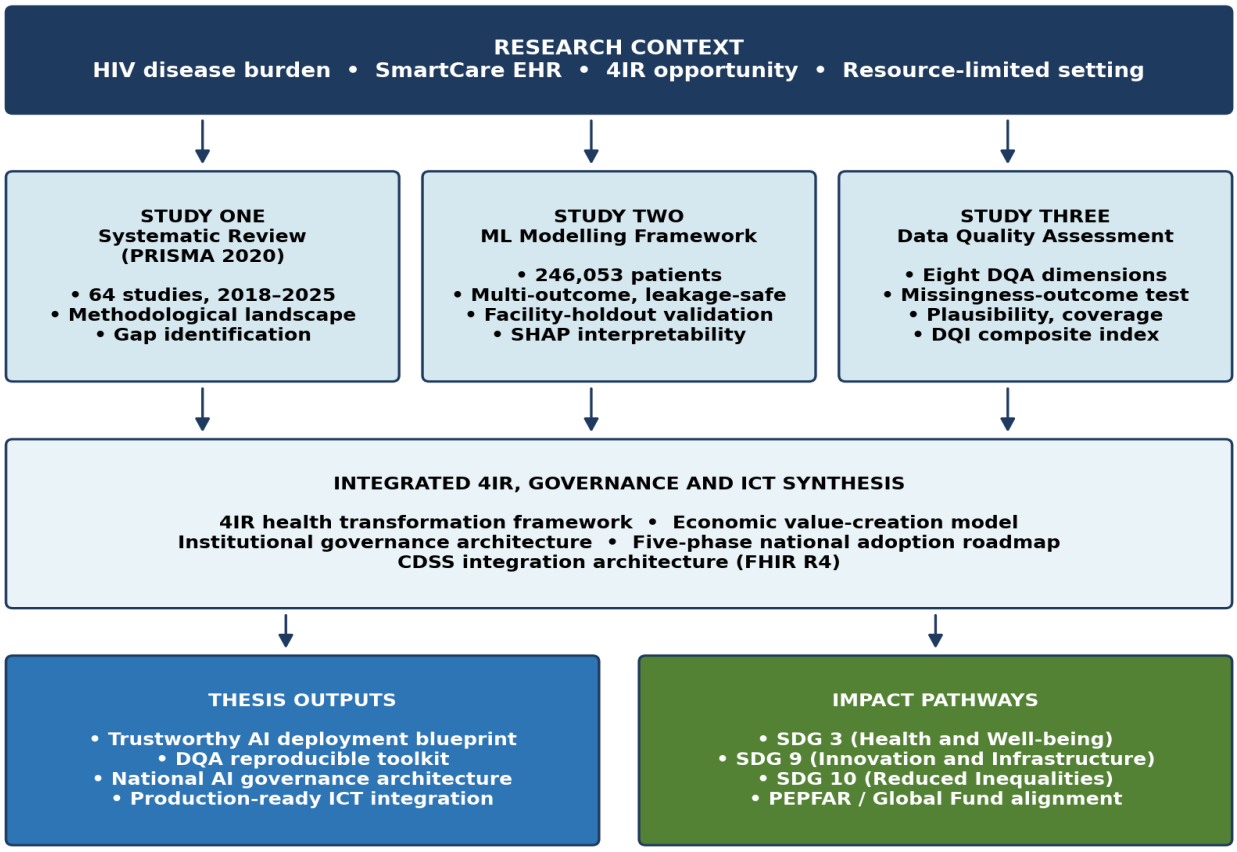

Figure 1.1: *Overall research framework integrating three empirical study components and the 4IR economic, governance and ICT synthesis into a coherent contribution to trustworthy, deployable AI in resource-limited public health.*

Study Three (data quality assessment) is positioned as the analytical foundation, because understanding the reliability and completeness of the input data is a prerequisite for interpreting the outputs of Study Two (the ML framework). Study One (systematic review) provides the evidential and methodological context that motivates and situates both empirical studies. The 4IR

framework and ICT architecture synthesise these empirical findings into a deployable, sustainable blueprint for national-scale AI adoption.

The conceptual framework operationalises three theoretical perspectives. The first is post-positivism in health informatics research, which acknowledges that systematic empirical inquiry can produce objective knowledge of health system phenomena while recognising the constraints imposed by measurement error and contextual factors [75]. The second is the socio-technical systems perspective on health information technology, which emphasises that the success of AI adoption depends as much on workflow integration, governance and trust as on the underlying algorithm. The third is the equity-and-distributive-justice perspective on AI in public health, which foregrounds the distributional consequences of algorithmic decisions and their implications for vulnerable populations [47], [53].

These three perspectives jointly shape the analytical choices made throughout the thesis. The post-positivist commitment motivates the rigorous quantitative evaluation design, including facility-holdout external validation and the missingness-outcome confounding test. The socio-technical perspective motivates the integrated treatment of governance, workforce capacity and ICT architecture alongside the methodological contributions. The equity-and-distributive-justice perspective motivates the explicit attention to algorithmic fairness, the human-in-the-loop review for missingness-driven flagging and the patient-representation provisions in the proposed governance committee.

1.8 Zambia's HIV Programme: Historical and Operational Context

The empirical work of this thesis is situated within Zambia's national HIV programme, which has scaled from the first recognised AIDS cases in the late 1980s to one of the strongest epidemic responses in sub-Saharan Africa, with a reported 98-98-97 cascade that exceeds the UNAIDS 95-95-95 targets [29]. This success now rests on a mature digital health ecosystem (SmartCare, DHIS2, DISA, ELMIS and DATIM) and on a differentiated service-delivery model that allocates care intensity by patient risk. It is the residual gap behind these headline achievements, the minority of patients at risk of treatment interruption, virological failure or TB co-infection, that the predictive framework developed in this thesis is designed to address. The full epidemiological, programmatic and policy context, including the care cascade, the operational decision points and

the national 4IR and digital-health policy environment, is reviewed in Chapter 2 (Section 2.12); this section provides only the orienting summary needed to motivate the problem statement that follows.

1.9 Organisation of the Thesis

The thesis is organised into six chapters, supported by a comprehensive references list and six appendices.

Chapter 1 (Introduction) presents the research context, the 4IR opportunity, the problem statement, objectives, research questions and conceptual framework, together with the scope and significance of the research.

Chapter 2 (Literature Review) synthesises the evidence base on ML methods, evaluation practices, data quality, governance, 4IR health transformation and digital health infrastructure, with LMIC settings as the primary focus. The chapter is structured around six thematic domains, identifies four cross-cutting gaps and presents the conceptual framework that motivates the empirical chapters.

Chapter 3 (Research Methodology) describes the research design, study setting, data sources, feature engineering, ML model architecture, validation strategy, explainability methods, DQA framework, ethical framework and proposed ICT architecture for clinical deployment. The chapter provides sufficient methodological detail to support independent reproduction of all empirical studies.

Chapter 5 (Proposed Framework and Presentation of Findings) presents the findings of all empirical studies in full, following the same Study One → Study Two → Study Three sequence used throughout the thesis. Section 5.2 reports the systematic review findings (Study One); Sections 5.3 and 5.4 report the ML model performance and SHAP interpretability findings (Study Two); Section 5.5 reports the data quality assessment findings (Study Three), which explain and qualify the ML results by characterising the reliability of the data on which they depend; and Section 5.6 presents the cross-study synthesis matrix that links findings across the studies.

Chapter 6 (Discussion of Findings) interprets the findings across all studies, contextualises them within the prior literature, and synthesises their implications for practice, policy, 4IR adoption and

the field. The chapter integrates evidence from the companion economic and ICT architecture analyses, presents the 4IR health transformation framework, the institutional governance architecture, the five-phase adoption roadmap and the CDSS integration architecture, and incorporates feedback from the three stakeholder dissemination meetings.

Chapter 7 (Conclusions and Recommendations) summarises the contributions across the methodological, empirical, governance, economic and ICT dimensions, provides sixteen actionable recommendations organised by stakeholder group, including the five-phase AI adoption roadmap and governance architecture, acknowledges limitations and outlines directions for future research.

References provides full IEEE-format citations for the literature drawn upon throughout the thesis.

Appendices A through F provide the ethical approval and research authorisation documentation, the systematic review protocol and PRISMA checklist, representative SHAP waterfall plots, the full facility-level DQA detail tables, the reproducible code and configuration inventory and the minutes of the two dissemination meetings with the stakeholder feedback synthesis.

1.10 Chapter Summary

This chapter has established the research context, motivation and structure of the thesis. The convergence of expanded digital health infrastructure in Zambia, the maturation of methodological frameworks for trustworthy AI and the institutional readiness signalled by the Data Protection Act, the National Health Strategic Plan and active Ministry of Health engagement, creates an opportunity to demonstrate responsible AI deployment in resource-limited public health. Four interlinked gaps in the existing literature — methodological rigour, data quality, multi-outcome external validation and ICT deployment architecture — motivate the integrated four-component contribution of this thesis.

The work is grounded empirically in a deduplicated analytical cohort of 246,053 PLHIV across six Lusaka District facilities, draws methodologically on the PRISMA 2020 framework and the PROBAST-AI risk-of-bias instrument, and articulates a conceptual framework that integrates post-positivism, socio-technical systems thinking and the equity-and-distributive-justice perspective. The remainder of the thesis develops this contribution through systematic review,

empirical modelling, data quality assessment, and governance and ICT architecture design. The next chapter reviews the prior literature that situates and motivates these contributions.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter synthesises the existing literature on the application of AI and ML to EHR-based predictive analytics in public health, with particular focus on LMIC settings. The review is structured around six thematic domains: electronic health records in resource-limited settings; algorithmic methods and their practical suitability for LMIC contexts; data quality, preprocessing and feature engineering; model evaluation, validation and calibration practices; explainability, fairness and ethics; and the comparative critical synthesis of related works. The review is grounded in the PRISMA 2020 framework [20] and applies an adapted PROBAST risk-of-bias instrument extended to ML-specific considerations [21], [22].

The central argument of this chapter is that methodological investment in algorithmic novelty has substantially outpaced investment in evaluation rigour, explainability, governance and institutional readiness, and that this imbalance is most consequential precisely in the LMIC settings that carry the greatest burden of preventable disease and have the most to gain from responsible 4IR adoption. The chapter is organised so that each thematic section presents its own literature in narrative form, with cross-cutting analytical syntheses in Sections 2.7 (Critical Review of Related Works), 2.8 (Comparison with Related Works), 2.9 (Identified Gaps), 2.10 (Conceptual and Theoretical Framework) and 2.11 (Proposed Framework Components). Section 2.12 provides the chapter summary.

Reading conventions for the chapter are as follows. Where a source is repeatedly cited across thematic sections, the first citation provides the full contextual framing and subsequent citations refer back to the earlier discussion using the section reference convention "see Section 2.x". This convention is used deliberately to minimise textual repetition. Where two sections genuinely require parallel treatment of a single source — for instance, when a study contributes both methodologically and ethically — the second treatment focuses on the dimension not previously discussed.

2.2 Electronic Health Records in Resource-Limited Settings

2.2.1 EHR Infrastructure in Sub-Saharan Africa

Health systems in LMICs typically comprise community health posts, primary health centres, district hospitals and national referral facilities operating under severe resource constraints [22], [23]. EHR infrastructure is correspondingly heterogeneous. Many primary health care facilities in sub-Saharan Africa continue to use paper or hybrid records [24]. Where electronic systems exist, common platforms include OpenMRS (widely used in East and Southern Africa), DHIS2 (for aggregate national reporting), iSantePlus (Haiti and Caribbean) and SmartCare (Zambia's national HIV EHR platform) [3], [24]. Each platform reflects design choices appropriate to its initial deployment context, and each has been adapted by national programmes in ways that produce facility-level schema heterogeneity and cross-platform interoperability challenges.

Such systems are designed primarily for programme reporting rather than research-grade longitudinal data collection. The consequences for data quality are well documented. First, systematic missingness arises from patients accessing care at multiple facilities without record linkage; in Zambia, an estimated 8–15% of PLHIV transfer between facilities at least once in their treatment journey, with biometric identifiers helping but not eliminating duplication [4], [13]. Second, inconsistent coding of diagnoses and laboratory results produces schema heterogeneity even within a single national platform; the data extraction pipeline used in this thesis required 23 separate conditional checks to handle fields that were present in some facilities but absent or differently named in others. Third, limited longitudinal depth occurs because patient mobility and re-enrolment generate fragmentary records of variable duration. Fourth, incomplete digitisation of historical paper records leaves the pre-digital era effectively invisible to AI training [16], [25].

Infrastructure barriers compound these data quality challenges. Unreliable electricity, limited internet connectivity and insufficient computational resources restrict the deployment of data-intensive models [26], [27]. Zambia's 2025 data confirm that only 883 of 3,540 health facilities (25%) have internet connectivity, a critical constraint for any national AI deployment [28]. The implication is that any AI architecture designed for national scale must be offline-resilient by default, not as an optional add-on.

2.2.2 Zambia's SmartCare HIV EHR System and Digital Health Ecosystem

SmartCare is Zambia's national HIV electronic health record platform, in operation since 2004 and deployed across more than 1,600 high-volume public health facilities nationwide, with over two million patients enrolled [3], [4]. The system captures patient demographics, enrolment details, visit and clinical event records, programme status and outcome indicators and laboratory results. It is the primary data infrastructure for Zambia's HIV treatment programme and the foundation for Ministry of Health reporting to PEPFAR and UNAIDS. Zambia has achieved notable programmatic progress: 98% of PLHIV are aware of their status, 98% of those diagnosed are receiving antiretroviral therapy (ART), and 97% of those on ART are virally suppressed, exceeding the UNAIDS 95-95-95 targets [29]. In FY2024, 1,295,030 PLHIV were receiving life-saving ART through this system [29].

Zambia's digital health infrastructure extends beyond SmartCare. DHIS2 provides aggregate national reporting across approximately 90% of health facilities. DISA manages laboratory information, including viral load and CD4 test results. ELMIS manages pharmaceutical supply chain data. DATIM supports PEPFAR reporting. Collectively, these interconnected systems generate a comprehensive digital footprint of the HIV programme at facility, district, provincial and national levels [7], [8]. This ecosystem provides the foundation for AI-driven clinical decision support, but also raises significant integration, interoperability and data quality challenges that any deployment architecture must address.

Despite its extensive coverage, SmartCare presents several data quality challenges directly relevant to AI applications. Data entry is frequently performed retrospectively by data clerks working from paper records; field completeness is highly variable across facilities and time periods; and the same information field may be named differently across facilities, creating schema heterogeneity [3], [13]. These characteristics make SmartCare a representative example of the EHR data quality challenges faced by HIV programmes across sub-Saharan Africa, and the primary empirical focus of this thesis.

The architectural evolution of SmartCare is itself instructive. The platform began as a single-facility client-server system, was progressively scaled with PEPFAR and Centers for Disease Control and Prevention (CDC) support, and now operates as a distributed national system with facility-level instances synchronising to a national data warehouse [3]. The current architecture

exposes data to research and analytics through scheduled bulk CSV exports rather than a native FHIR API; this is a critical constraint for the proposed CDSS integration architecture discussed in Chapter 6, which therefore requires a FHIR Adapter Service to translate CSV exports into HL7 FHIR R4-compliant resources.

2.2.3 The Fourth Industrial Revolution and Health System Transformation

Schwab's conceptualisation of the Fourth Industrial Revolution identifies three technology clusters driving transformation: physical systems (autonomous vehicles, advanced robotics); digital systems (AI, big data, IoT); and biological systems (genomics, synthetic biology) [1]. In health systems, the most immediately consequential cluster for LMICs is the digital one, and particularly AI and ML applied to health data [3], [4]. The digital cluster is consequential because it has the lowest infrastructure threshold for adoption: ML models can be developed on commodity hardware, deployed on conventional server infrastructure, and integrated into existing EHR systems without prohibitive capital investment, provided that the data foundations are adequate.

For LMICs, the most tractable and impactful 4IR health technology is not necessarily the most algorithmically advanced. Evidence consistently shows that, in resource-constrained settings with fragmented data infrastructure, interpretable ML models such as logistic regression and gradient boosting outperform complex deep learning architectures in practical deployment, because they require less data, less compute and produce more clinically transparent outputs [15], [43]. This has important economic implications: LMIC health systems do not need to invest in cutting-edge AI infrastructure to realise the value of 4IR analytics. They need to invest in the data quality foundations and governance structures that make any ML approach reliable and equitable.

Several African countries have leveraged 4IR technologies for health system improvement. Rwanda's use of drone technology for blood-product delivery has reduced last-mile distribution times and improved availability of life-saving products [17]. South Africa's deployment of ML for tuberculosis diagnosis from chest X-rays has demonstrated near-radiologist performance for triage [17], [18]. Kenya's integration of mobile health platforms into community health systems has enabled real-time data capture and feedback loops at sub-county level [18]. These examples show that 4IR health adoption in Africa is not hypothetical, but they also illustrate a consistent pattern: successful adoption is preceded by substantial investment in foundational infrastructure (connectivity, digital literacy, system integration) and governance (regulatory frameworks, data

protection laws, ethics oversight) [5], [17]. Zambia's HIV programme represents one of the most mature digital health contexts in which 4IR analytics can be applied responsibly, and the present thesis is an attempt to demonstrate that responsibility in evidence and design.

2.3 Machine Learning for Public Health Prediction

2.3.1 Overview of ML Method Categories

ML methods learn patterns from data to make predictions. In EHR-based health research, the goal is to build a model that maps patient information (demographics, diagnoses, laboratory results, visit history) to a health outcome of interest. EHR datasets present specific challenges that distinguish them from the curated benchmark datasets on which ML methods are typically developed: high dimensionality, missing data, imbalanced outcomes, irregular time intervals between visits, distributional shift over time, and frequent presence of free-text data that requires natural language processing [29], [30]. The breadth of these challenges has led to a corresponding breadth of method development across four broad categories.

Traditional ML methods, principally logistic regression and decision trees, offer transparent outputs, low computational requirements and reliable calibration even with sparse data, making them well suited to LMIC environments [31], [32]. Their advantages are that they are well-understood, easy to inspect and debug, and supported by mature reporting standards such as TRIPOD [71]. Their main limitation is that they capture linear or simple non-linear relationships and may miss higher-order interactions that are clinically meaningful.

Ensemble methods, including random forests, gradient boosting and XGBoost, handle noisy or incomplete data robustly, generate feature importance scores that support model interpretation and typically outperform single traditional models on tabular EHR data [33], [34]. Random forests reduce overfitting by averaging predictions across many independent decision trees [33]. Gradient boosting methods such as XGBoost build trees sequentially, with each tree correcting the residuals of the previous tree, and have repeatedly won predictive modelling competitions on tabular data [34]. The stacked ensemble architecture used in this thesis (Chapter 3) combines logistic regression, random forest and XGBoost into a meta-learner, exploiting the complementary strengths of each base model.

Deep learning architectures, including feedforward networks, recurrent neural networks (LSTMs) and transformers, can capture complex temporal patterns in longitudinal EHR data [35]–[37]. The Rajkomar et al. study of scalable deep learning on EHRs in two US academic health systems used over 46 billion data points and demonstrated near-state-of-the-art performance across multiple prediction tasks [5]. The MIMIC-III critical care database has supported a flowering of deep learning applications [30]. However, deep learning approaches require large, well-curated datasets and substantial computational infrastructure, generally limiting their utility to high-resource settings. The 246,053-patient cohort used in this thesis is large by LMIC standards but does not approach the scale of the MIMIC-III or Rajkomar datasets, and the cohort has the additional challenge that engagement and laboratory features are computable for only 12.3% and 3.3% of patients respectively. These constraints favour traditional ML and ensemble methods over deep learning for the Zambian context.

Hybrid and emerging approaches, including federated learning and AutoML, offer potential advantages in privacy preservation and automation [38], [39]. Federated learning allows model training across multiple sites without sharing raw patient data, which is potentially attractive for national or multi-country networks where data sharing is restricted by privacy regulation [38]. AutoML approaches automate hyperparameter search and model selection, reducing the technical barrier to ML development [39]. These approaches are not deployed in the empirical work of this thesis, but they are discussed as candidates for future extension in Chapter 7.

2.3.2 Class Imbalance and Calibration

Two issues are especially important for public health settings, and both arise prominently in the empirical work of this thesis. First, class imbalance is the norm rather than the exception in HIV care prediction. The treatment interruption outcome (IIT12m) in the present cohort has a prevalence of 0.63% in the held-out test set; the unsuppressed viral load outcome (UVL12m) has a prevalence of 1.17%. Models trained without explicit attention to class imbalance can appear accurate simply by predicting the majority class. Class imbalance is addressed in this thesis through weighted loss functions (class_weight="balanced" in logistic regression and random forest; scale_pos_weight set to the ratio of negatives to positives in XGBoost) and through F1-optimised decision threshold tuning rather than the default 0.5 [40].

Second, calibration is the property that predicted probabilities correspond to real-world event rates. A model that predicts 30% risk for a group of patients is well-calibrated when 30% of that group go on to experience the event, and poorly calibrated when the actual rate is, say, 10% or 60%. Calibration is critical for triage and resource allocation decisions, where the prediction is used to allocate scarce intensive interventions to the highest-risk patients [41], [42]. The Brier score, the integrated calibration index and calibration plots are standard assessment tools, and the reporting standard TRIPOD-AI requires calibration as a mandatory reporting element [72]. Yet, as Section 2.5 documents, calibration is reported in only 6.2% of the included studies in the systematic review (Chapter 5).

2.3.3 Comparative Performance in LMIC Settings

A consistent finding across the literature is that performance gains from algorithmic complexity are often modest, particularly where data quality is poor or sample sizes are small [43]. The Christodoulou meta-analysis compared ML methods with logistic regression across multiple clinical prediction tasks and found no consistent advantage of ML over logistic regression [15], [43]. This finding is particularly relevant where data sparsity constrains complex architectures [28], [44]. Among LMIC-focused studies included in the systematic review of this thesis, traditional ML and ensemble methods are each dominant (33.3% each), reflecting pragmatic alignment between model choice and the data quality, interpretability and infrastructure constraints of resource-limited health systems [24], [28].

Specific LMIC studies illustrate the practical patterns. Mutai et al. used logistic regression and random forest to predict HIV risk in Kenya using routine clinic data, achieving AUC-ROC around 0.75 for a single outcome [49]. Musukwa et al. used logistic regression and XGBoost on SmartCare data to predict treatment outcomes in Zambia, achieving moderate discrimination but reporting no calibration or external validation [50]. Gichuki and Maina applied a range of ML methods to predict HIV treatment interruption in Kenya [48]. Kiyingi et al. used random forest and gradient boosting to predict loss to follow-up in Uganda [64]. Karimanzira et al. studied TB risk prediction in PLHIV using SmartCare data in Zambia [65]. Krasniuk et al. studied treatment failure prediction in Zambia [66]. Collectively, these studies demonstrate the feasibility of ML prediction in LMIC HIV care, but persistent gaps remain in external validation, calibration assessment, multi-outcome

integration and SHAP-based interpretability — gaps which the present thesis is designed to address.

2.3.4 HIV-Specific Prediction Tasks

The three outcomes addressed in this thesis — recorded TB treatment (TB12m), programme-recorded interruption in treatment (IIT12m) and recorded unsuppressed viral load (UVL12m) — each have established clinical and operational rationale and a corresponding methodological literature. TB co-infection is the leading cause of HIV-related mortality globally, and PLHIV face a 20- to 30-fold higher risk of active TB than the general population [3]. Predictive identification of patients at risk of TB incidence supports targeted intensive screening, isoniazid preventive therapy and earlier active case finding [3], [65]. Treatment interruption is the most operationally consequential threat to programme performance, undermining viral suppression at population level and contributing to onward transmission; predictive identification supports targeted retention interventions through differentiated service delivery models [19], [48], [64]. Unsuppressed viral load identifies patients with active virological failure, who require enhanced adherence counselling, resistance testing and potentially regimen optimisation [66].

A specific methodological consideration arises for the TB12m outcome. The 36% overall TB prevalence reported in the empirical cohort substantially exceeds typical TB incidence rates in PLHIV cohorts in the published literature [3], [65]. The companion DQA paper documents that this elevated prevalence reflects label conflation: the available data fields do not reliably distinguish between prevalent TB (already present at index date) and incident TB (newly diagnosed within the 12-month follow-up window), so the operationalised TB12m outcome combines both. This is itself an important DQA finding (D5: Outcome Sensitivity, Chapter 3) and is discussed at length in Chapter 6.

2.3.5 Ensemble Learning Strategies: A Critical Comparison of Bagging, Boosting and Stacking

Because the empirical framework developed in this thesis is an ensemble, the three dominant ensemble paradigms warrant explicit comparison rather than assumed equivalence. They differ not merely in implementation but in the error component each is designed to reduce, and that difference determines which is appropriate for sparse, imbalanced, heterogeneous EHR data.

Bagging (bootstrap aggregating) trains base learners in parallel on bootstrap resamples and averages their outputs [150]. Because the resamples are independent, bagging principally reduces variance and leaves bias largely unchanged; random forests extend this by additionally decorrelating trees through random feature subsampling at each split [33]. The practical consequence is robustness: bagged ensembles are stable under noisy predictors and rarely overfit as the number of trees grows, but they cannot recover signal that individual weak learners systematically miss, and they are relatively insensitive to rare positive classes.

Boosting fits base learners sequentially, each new learner weighting the residual errors of its predecessors, and therefore principally reduces bias. Gradient-boosted decision trees, of which XGBoost [34] is the most widely used implementation, dominate applied tabular prediction; LightGBM [143] achieves comparable accuracy at lower computational cost through histogram-based, leaf-wise growth, and CatBoost [144] introduces ordered boosting and native categorical handling that mitigate target leakage in high-cardinality categorical data. The corresponding weakness is sensitivity: boosting can overfit noise and mislabelled outcomes, which is a material risk where outcome ascertainment depends on incomplete routine documentation, as it does in this study.

Stacking, or stacked generalisation [81], differs from both in that it does not aggregate by a fixed rule. Base learners are trained on the data and a meta-learner is trained on their out-of-fold predictions, so the combination rule is itself learned from data. Stacking can therefore exploit complementarity: where a linear learner captures monotone marginal effects, a bagged learner captures interaction-driven stability and a boosted learner captures sharp non-linear structure, the meta-learner can weight these contributions per-outcome rather than uniformly. Its costs are additional variance in the meta-level fit, greater computational expense, and reduced transparency, and it yields little benefit when base learners are highly correlated in their errors.

Table 2.1 summarises this comparison. The relevance to the present study is direct: the three outcomes modelled here differ markedly in prevalence and in the structure of their predictive signal, so a fixed aggregation rule optimised for one outcome is unlikely to be optimal for the others. This is the specific methodological reason a learned combination was selected over bagging or boosting alone, and it is the hypothesis that the model-comparison and significance testing reported in Chapter 5 are designed to test.

Table 2.1: Critical comparison of ensemble learning strategies and their suitability for sparse, imbalanced routine EHR data.

Strategy	Mechanism	Error component reduced	Principal strengths	Principal weaknesses	Suitability for this study
Bagging (incl. random forest) [150], [33]	Parallel training on bootstrap resamples; averaged output	Variance	Stable under noisy predictors; resistant to overfitting; native importance measures	Cannot correct systematic base-learner bias; relatively insensitive to rare positives	Useful as a stabilising component, insufficient alone for rare outcomes (IIT12m 0.53%)
Boosting (XGBoost, LightGBM, CatBoost) [34], [143], [144]	Sequential fitting to predecessor residuals	Bias	Strong accuracy on tabular data; handles sparsity and imbalance weighting	Sensitive to label noise and outcome misclassification; more tuning-dependent	Strong single learner, but exposed to outcome-ascertainment noise in routine records
Stacking [81]	Meta-learner trained on out-of-fold base predictions	Bias and variance jointly, via learned weighting	Exploits complementary error structures; adapts weighting per outcome	Added meta-level variance; higher cost; lower transparency	Selected: permits one pipeline to adapt across three outcomes of differing prevalence

2.3.6 Sequential Deep Learning for EHR Time Series: Recurrent and Transformer Architectures

The preceding subsections treat prediction as a tabular problem, in which each patient is represented by a fixed-length feature vector constructed at an index date. A substantial and rapidly growing literature instead treats the electronic health record as a sequence of timestamped events and applies architectures designed for temporal data. Because this literature bears directly on the modelling choices made in this thesis, it is reviewed here explicitly rather than left implicit.

Recurrent architectures were the first to be applied at scale. Long short-term memory (LSTM) and gated recurrent unit (GRU) networks maintain a hidden state across successive clinical events and thereby model order and duration effects that a static feature vector discards. A comprehensive review of deep learning for healthcare time series concludes that LSTM and GRU perform comparably in most clinical applications, that GRU is the less complex and faster of the two, and that bidirectional variants outperform unidirectional ones for long event sequences [148]. Rajkomar et al. demonstrated the scalability of this approach across multiple prediction tasks using raw, unharmonised EHR data [5], and reviews of deep learning in healthcare have consolidated the architectural landscape [57].

Transformer architectures subsequently displaced recurrent models as the dominant sequential approach. BEHRT [146] and Med-BERT [147] adapt the bidirectional transformer to structured

EHR by treating diagnosis and visit codes as tokens with learned positional embeddings, enabling transfer of a pre-trained representation to multiple downstream prediction tasks. More recently, large language models have been applied to structured EHR either by serialising tabular records into text or by using their embeddings as encoders, and have been shown to improve transferability of EHR-based predictions across countries and coding systems [157]; a systematic review of generative LLM applications to real-world EHR data identified 196 qualifying studies, of which clinical decision support constituted the largest share [158], while task-specific models such as DeLLiriuM demonstrate structured-EHR outcome prediction directly [159].

The evidence most directly relevant to this thesis comes from HIV care in sub-Saharan Africa. Mirugwe et al. [149] compared non-sequential learners with a transformer (BERT) model for visit-level prediction of missed HIV appointments using UgandaEMR data from 66,206 people living with HIV, 1,479,121 clinical visits and 86 facilities, and reported an AUC of 0.96 (95% CI 0.94–0.97) for the transformer against 0.90 (95% CI 0.87–0.92) for the best-performing gradient-boosted alternative. This result establishes that sequential architectures can add material value on longitudinal African HIV programme data and that a blanket dismissal of deep learning in this setting would be unwarranted.

Three qualifications, however, govern how that finding transfers to the present study, and they are stated here because they motivate the design adopted in Chapter 4. First, the prediction targets differ substantially: the Ugandan model predicts attendance at the next scheduled visit, a short-horizon, visit-level outcome with a prevalence of 10.7%, whereas this thesis predicts three twelve-month patient-level outcomes with prevalences as low as 0.53%, a regime in which sequence length per patient is shorter and the positive class far rarer. Second, the comparison in that study was, as its authors explicitly acknowledge, not conducted under identical feature constraints: the transformer received ordered visit sequences while the tabular learners received static representations, so the reported margin reflects a difference in input representation as much as in architecture. Third, records with missing values were excluded, removing 37.6% of clients, which the authors note may bias the analytic sample toward better-documented patients and overstate attainable performance — precisely the failure mode that the data-quality assessment in this thesis is designed to expose.

Set against this, the benchmark literature on tabular prediction consistently finds that gradient-boosted tree ensembles remain highly competitive with, and frequently superior to, deep architectures on structured data of moderate dimensionality, particularly where features are heterogeneous, uninformative predictors are numerous and sample sizes are not extreme [141], [142], [145]. Cross-representation benchmarking on longitudinal EHR further indicates that sequential models are constrained by feature sparsity and irregular sampling — the dominant characteristics of the dataset analysed here, where engagement-feature coverage is 12.3%.

The position taken in this thesis is therefore neither that deep learning is unnecessary nor that it is uniformly superior, but that under the specific conditions of this study — extreme class rarity, sparse and irregularly sampled engagement data, a documented Data Quality Index below 70%, an explainability requirement for clinical governance, and a CPU-only, offline-resilient deployment constraint — a leakage-safe tabular ensemble is the defensible primary design, while sequential architectures constitute the principal avenue for extension. Their formal evaluation against the present models is reported as a comparative experiment in Chapter 5 and identified as a priority for future work in Chapter 7.

2.3.7 Recent Research Closely Related to This Study (2024–2026)

Finally, and in order to position the thesis against the current state of the field rather than against the literature available at the time the study was designed, this subsection consolidates recent closely-related research. Four strands are salient. The first is the continued expansion of machine learning for HIV retention and treatment-interruption prediction across sub-Saharan Africa, which establishes both the programmatic demand for such models and the persistence of the evaluation gaps quantified in Study One. The second is the application of large language and foundation models to structured EHR prediction, reviewed in Section 2.3.6, whose principal contribution to date has been transferability and representation rather than demonstrated clinical effectiveness in low-resource settings. The third is methodological guidance for the reporting and appraisal of AI prediction models, notably the TRIPOD-AI and PROBAST-AI extensions [72], which formalise the external-validation and calibration expectations that this thesis adopts. The fourth is the equity and governance literature, including frameworks for equity-disaggregated performance monitoring of clinical AI [139] and stakeholder analyses of AI deployment in low- and middle-income countries [127].

Table 2.2 summarises representative recent studies against the dimensions on which this thesis makes its contribution. Two patterns are evident across the recent literature. External and prospective validation remain the exception rather than the rule, so the asymmetries identified in Study One have not been resolved by the most recent work. Equally, data-quality certification is almost never treated as a precondition of model development, notwithstanding that several of these studies report substantial record exclusion arising from incompleteness. The integration proposed in this thesis therefore remains unaddressed in the current literature.

Table 2.2: Summary of recent (2024–2026) research closely related to this study, and its position relative to the present work.

Study	Setting and data	Method	Principal finding	Gap relative to this thesis
Mirugwe et al. (2026) [149]	Uganda; UgandaEMR; 66,206 PLHIV; 1,479,121 visits; 86 facilities	Transformer (BERT) vs decision tree, random forest, AdaBoost, XGBoost	Transformer AUC 0.96 (0.94–0.97) vs 0.90 (0.87–0.92) for best tabular model, visit-level missed appointments	No external validation; 37.6% of clients excluded for missingness; no fairness or data-quality certification
Ngcobo et al. (2025) [55]	Sub-Saharan Africa; systematic review	Review of AI applications to HIV care	Rapid growth in AI-for-HIV research with persistent reporting weaknesses	Confirms evaluation gaps; provides no deployable framework or DQA instrument
Kirchler et al. (2026) [157]	Multi-country EHR; cross-coding-system evaluation	LLM embeddings with transformer prediction head	LLM representations improve cross-country and cross-coding transferability	High-income data and infrastructure; not evaluated under CPU-only, low-connectivity constraints
Du et al. (2026) [158]	Systematic review of 196 studies of generative LLMs on real-world EHR	PRISMA-guided systematic review	Clinical decision support is the largest application class; evaluation practice is heterogeneous	Documents evaluation heterogeneity rather than resolving it; LMIC representation minimal
Moons et al. (2024) [72]	Methodological guidance	TRIPOD-AI / PROBAST-AI reporting and appraisal extensions	Formalises external validation, calibration and fairness reporting expectations	Guidance only; this thesis operationalises these requirements empirically in an LMIC HIV programme
Liu et al. (2025) [139]	Clinical AI monitoring	Equity-disaggregated performance monitoring framework	Subgroup-disaggregated monitoring is necessary for equitable clinical AI	Framework not instantiated on LMIC HIV data; this thesis reports measured subgroup fairness (Section 5.12)
Wang et al. (2024) [113]	Routine HIV care data	Multilevel modelling of structural and clinical drivers	Structural determinants materially shape HIV outcomes in routine care	Explanatory rather than predictive; no deployment or governance pathway
This thesis (2026)	Zambia; SmartCare; 246,053 PLHIV; six Lusaka facilities	Leakage-safe multi-outcome stacked ensemble; facility-holdout validation; eight-dimension DQA; governance and ICT architecture	Multi-outcome prediction (AUC 0.707–0.839) certified against an explicit data-quality gate and translated to a deployment architecture	—

2.4 Data Quality and Preprocessing in EHR-Based ML

2.4.1 Data Quality Frameworks for Health Data

Data quality challenges are a pervasive theme in the prior literature, particularly for LMIC contexts. The seminal framework of Weiskopf and Weng [16] identifies five dimensions of EHR data quality for clinical research reuse: completeness, correctness, concordance, plausibility and currency. This framework has been widely cited and underpins numerous subsequent data quality assessment instruments. However, the Weiskopf and Weng framework was developed for high-income clinical research settings, and as Dong et al. note [68], it does not directly address several dimensions that are essential for AI training rather than for traditional clinical research, including feature coverage (the proportion of patients for whom each feature is computable), outcome sensitivity (the stability of outcome labels across operationalisations) and missingness-outcome confounding (the systematic association between data completeness patterns and outcome rates).

The eight-dimension DQA framework developed in this thesis (Chapter 3, Section 3.5) extends the Weiskopf and Weng framework with three AI-specific dimensions: feature coverage, outcome sensitivity and missingness-outcome confounding. These three additions are operationalised quantitatively in Chapter 5, with results that reveal the substantial gap between traditional data quality assessment and AI-readiness assessment. The companion DQA paper, preprinted under DOI 10.21203/rs.3.rs-9306862/v1, develops the framework in full.

Common issues documented across the prior literature include systematic missingness, inconsistent diagnosis coding, fragmented patient records across facilities and limited longitudinal depth [16], [25]. A recurring methodological weakness is that preprocessing strategies, including imputation method choice, temporal aggregation windows and feature selection criteria, are rarely evaluated through sensitivity analyses. This makes it difficult to attribute reported model performance to architecture rather than data preparation decisions [67]–[69]. The leakage-safe temporal feature engineering pipeline used in this thesis (Chapter 3) is designed to be transparent and reproducible, with all preprocessing parameters fitted exclusively on the training set and explicitly documented for independent re-implementation.

2.4.2 Missingness Mechanisms

A useful classification distinguishes three missingness mechanisms with different statistical and clinical implications [70]. Missing completely at random (MCAR) occurs when the probability of missingness is independent of both observed and unobserved data; under MCAR, simple imputation and complete-case analysis produce unbiased estimates, but MCAR is rarely a realistic assumption in routine EHR data. Missing at random (MAR) occurs when the probability of missingness depends on observed data but not on unobserved data; under MAR, model-based multiple imputation (such as MICE) can produce unbiased estimates [25], [70]. Missing not at random (MNAR) occurs when the probability of missingness depends on the unobserved value itself; for example, sicker patients may miss visits more often, so the missingness of an outcome measurement may correlate with the value of the unobserved measurement. MNAR is the most challenging case and the most realistic for HIV EHR data, where patients who interrupt treatment by definition stop generating clinical records.

The missingness-outcome confounding test introduced as D8 in the DQA framework (Chapter 3) is designed to surface MNAR patterns at the population level. By computing Pearson correlations between facility-level missingness rates and outcome rates, the test identifies feature-outcome pairs where the missingness pattern is systematically associated with the outcome, indicating that AI predictions trained on the data may be partly driven by the missingness pattern rather than by the clinical signal. The empirical findings (Chapter 5) show three feature-outcome pairs with $|r| > 0.5$ in the SmartCare cohort: education-TB (+0.599), education-UVL (−0.865) and stage-TB (+0.741), each with substantive implications discussed in Chapter 6.

2.4.3 Imputation Strategies

Imputation strategies for EHR data include simple imputation (mean, median, most-frequent), regression-based imputation, multiple imputation by chained equations (MICE), k-nearest neighbours imputation and indicator-based "missingness flag" imputation [25], [70]. Each has well-documented strengths and weaknesses, and the choice of imputation strategy can substantially affect downstream model performance [70]. The thesis adopts a deliberately conservative imputation strategy: median imputation for numeric variables and most-frequent imputation for categorical variables, both fitted exclusively on the training set. This choice is conservative in the sense that it does not impose distributional assumptions and does not propagate

uncertainty from the imputation step into the model, but it has the corresponding limitation that the model is not informed about the uncertainty associated with imputed values.

A specific implementation feature of the preprocessing pipeline merits discussion. For categorical variables with missing values, the imputation strategy creates an explicit one-hot encoded "None" category, so that missingness itself becomes a feature visible to the model. This design choice is consequential: it allows the model to detect missingness patterns and to assign predictive weight to them, with both desirable and undesirable consequences. On the one hand, it surfaces the missingness-outcome confounding documented in Chapter 5, making the data quality issue visible rather than hidden. On the other hand, it allows the model to base predictions on missingness, which raises the equity risk discussed extensively in Chapter 6: AI tools deployed in routine clinical care could systematically flag patients whose education data is missing rather than patients with genuine clinical need.

2.4.4 Temporal Feature Engineering and Leakage Prevention

Information leakage — the inadvertent inclusion of post-index or outcome-related features in predictors — has been identified as a pervasive problem in clinical ML studies, inflating apparent performance and producing models that collapse on prospective deployment [14], [18]. Leakage takes many forms: outcome data used as features; future events available at the time of prediction; data preprocessing steps fitted on the combined train-test set; and time-varying features computed without an explicit temporal cut-off [14]. The leakage-safe pipeline used in this thesis enforces a strict temporal boundary: every feature is derived from records dated strictly before the patient's index date; outcomes are derived strictly from records dated strictly after the index date; and the strict equality of the index date is excluded from both predictor and outcome construction.

The Temporal API Gateway proposed in the CDSS integration architecture (Chapter 6) operationalises this temporal boundary at the system level rather than at the script level. Any feature extraction request that would return records dated after the patient's index date is automatically rejected by the gateway, with the rejection logged for audit. This design choice means that the leakage-safe property of the model is enforced architecturally and cannot be circumvented by individual code paths.

2.5 Evaluation, Validation and Calibration Practices

2.5.1 Discrimination Metrics

Evaluation practices in the AI/ML health literature have been widely critiqued for over-reliance on discrimination metrics, particularly AUC-ROC, at the expense of calibration and external validation [41], [43], [70]. AUC-ROC is the area under the receiver operating characteristic curve, which plots true positive rate against false positive rate across all possible decision thresholds [78]. AUC-ROC has the appealing property of being threshold-independent, but it has the corresponding limitation of being insensitive to class imbalance: a model can have a high AUC-ROC while assigning near-zero probability to the positive class for almost all patients [40], [79]. For the rare outcomes in this thesis (IIT12m at 0.63% prevalence; UVL12m at 1.17% prevalence), AUC-ROC is supplemented by AUC-PR, the area under the precision-recall curve, which is more informative for imbalanced data [40], [79].

At a specific operational threshold, additional metrics become relevant. F1 score is the harmonic mean of precision and recall, balancing the cost of false positives (precision) and false negatives (recall) at a specified threshold. Precision (positive predictive value) is the proportion of positive predictions that are correct, and recall (sensitivity) is the proportion of true positives that are correctly predicted. Specificity is the proportion of true negatives that are correctly predicted. Negative predictive value is the proportion of negative predictions that are correct. The selection of metrics for reporting should reflect the clinical decision being supported: high-sensitivity screening favours recall; intensive intervention triage favours precision; balanced operational deployment favours F1.

2.5.2 Calibration

Calibration, the agreement between predicted probabilities and observed outcomes, is essential for triage and resource allocation applications yet is systematically underreported [41]. Van Calster et al. argue that calibration is the "Achilles' heel" of predictive analytics, in that even models with strong discrimination can be poorly calibrated and therefore produce predicted probabilities that are not interpretable as real-world risks [41]. The Brier score, defined as the mean squared difference between predicted probability and observed outcome, is a standard summary metric; calibration plots and calibration-in-the-large statistics provide finer-grained assessment.

In the thesis, calibration is reported for the held-out test set predictions of each base model and the stacked ensemble. The systematic review (Chapter 5) confirms that calibration is reported in only 6.2% of the included studies, and only 4 of 64 studies report calibration with sufficient detail to permit cross-study comparison. The implication is that the operational reliability of published ML models in HIV care is largely unknown, even where discrimination metrics suggest strong performance. The thesis therefore makes calibration a mandatory reporting element in the model evaluation and a precondition for advancement to deployment in the proposed five-phase adoption roadmap (Chapter 6).

2.5.3 External Validation

External validation on independent datasets is rare in the published ML-in-LMIC literature [70], [71]. The systematic review identifies only 5 of 64 studies (7.8%) reporting external validation. The standard validation design in the field is the random patient-level train-test split, which divides the cohort into training and test subsets without regard to facility, time period or patient subgroup. This design tests how well a model generalises to new patients drawn from the same population as the training set, but it does not test how well the model generalises to new facilities, new time periods or new patient subgroups, which is the relevant condition for health system deployment.

The facility-holdout external validation design used in this thesis (Chapter 3) is a stricter test. Two complete facilities (Matero Main Urban Health Centre and Matero Reference Hospital, n = 74,506) are held out from training; the remaining four facilities (n = 171,547) are used for training only; and the model is evaluated exclusively on the two held-out facilities. This design tests cross-facility generalisability rather than within-facility performance, and is the relevant condition for deployment in health systems where the model is expected to be used at facilities not represented in the training data. The achievement of AUC-ROC of 0.839 for IIT12m under this stricter design (Chapter 5) is substantially stronger evidence of operational deployability than equivalent performance under a random patient-level split.

The TRIPOD-AI reporting standard, currently in development as an extension of the TRIPOD guideline for clinical prediction models [71], will likely make external validation a mandatory reporting element [72]. The thesis anticipates this standard and adopts facility-holdout external validation as the empirical default for cross-facility generalisability assessment.

2.5.4 Decision Threshold Selection

The default decision threshold of 0.5 is rarely optimal for clinical decision support, particularly for rare outcomes. F1-optimised threshold selection [40] computes the threshold that maximises F1 score on the precision-recall curve, producing operationally meaningful thresholds tuned to the prevalence and the clinical question. For the TB12m outcome in this thesis, the F1-optimised threshold is 0.257, supporting a high-sensitivity screen-and-refer strategy that identifies 92.3% of future TB cases. For the IIT12m and UVL12m outcomes, the F1-optimised thresholds are 0.902 and 0.928 respectively, reflecting the extreme class imbalance and identifying only the highest-certainty cases for intensive intervention. The thesis reports performance at both 0.5 and the F1-optimised threshold for full transparency.

An alternative threshold selection approach, not pursued in this thesis but discussed as future work in Chapter 7, is decision-curve analysis [14], which computes the net benefit of a prediction model at each decision threshold by weighing true positives against false positives at a clinician-specified harm ratio. Decision-curve analysis is particularly suited to clinical decisions where the costs of false positives and false negatives differ substantially, and is a natural extension of the F1-optimised approach used here.

2.6 Explainability, Fairness and Ethics in Health AI

2.6.1 Why Explainability Matters

The prior literature increasingly recognises that predictive performance alone is insufficient for responsible public health deployment of ML models. Clinicians, patients, regulators and governance committees all require some explanation of how a model arrived at its prediction in order to assess whether the prediction can be trusted and acted upon [52], [73], [74]. The absence of explainability in deployed black-box models is itself a deployment risk: clinicians may either over-trust the model (delegating clinical judgement to the algorithm) or under-trust the model (ignoring its recommendations regardless of accuracy), and both failure modes erode the clinical value of the system.

Explainability frameworks have therefore moved from optional to expected for any AI tool deployed in clinical decision support. SHAP (Shapley Additive Explanations) [52] provides a

mathematically principled approach derived from cooperative game theory, in which each feature's contribution to a specific prediction is computed as its marginal effect averaged across all possible feature subsets. SHAP explanations have the desirable property of additivity: for any individual prediction, the SHAP values of all features sum to the difference between the model prediction and the model base value. LIME (Local Interpretable Model-Agnostic Explanations) [73] is an alternative approach that fits a locally linear surrogate model around a single prediction point. Both approaches are widely used; SHAP has become dominant in the clinical ML literature owing to its theoretical guarantees and the availability of efficient implementations such as TreeSHAP for tree-based models.

2.6.2 Permutation Importance

Permutation importance is a complementary, model-agnostic interpretability method that quantifies the contribution of each feature to model performance by measuring the reduction in a chosen scoring metric when the feature's values are randomly permuted [33]. Unlike SHAP, which decomposes individual predictions, permutation importance is a population-level metric: it ranks features by their average contribution to model accuracy across the entire test set. The thesis reports both permutation importance and SHAP for each model and each outcome, with the comparison between the two metrics providing additional diagnostic insight into the model's feature use.

For the UVL12m outcome, permutation importance and SHAP global importance agree on the dominant feature (education level), which provides strong convergent evidence that this feature is genuinely driving the model's predictions. For TB12m and IIT12m, the two metrics rank the top features in different orders, with BMI dominant on permutation importance for TB12m and SHAP for IIT12m, and education dominant on permutation importance for IIT12m. The disagreement is itself informative, indicating that different features contribute differently to model accuracy at the population level (permutation importance) and to individual predictions (SHAP), reflecting the heterogeneity of clinical signal across the patient cohort.

2.6.3 Algorithmic Fairness and Equity

Ethical and governance frameworks, including WHO guidance [54] and the normative work of Vayena et al. [53], identify algorithmic bias, data sovereignty and transparency as priorities for LMIC contexts. The Obermeyer et al. study [47] demonstrated that health algorithms trained on

biased data can systematically disadvantage socioeconomically marginalised populations: a commercial algorithm used to allocate health resources in the United States underestimated the needs of Black patients because the proxy used to define need (historical health spending) reflected structural inequalities in access to care rather than underlying illness. This finding has direct relevance to the Zambian HIV care context, where systematic demographic missingness may cause AI tools to flag patients based on administrative data gaps rather than genuine clinical need [10], [11].

The fairness literature distinguishes several mathematical definitions of fairness, each with different implications [45], [46]. Demographic parity requires that the proportion of positive predictions is equal across subgroups. Equal opportunity requires that the recall (true positive rate) is equal across subgroups. Calibration parity requires that calibration is equal across subgroups. These definitions are mutually incompatible in general: a model that satisfies demographic parity may violate equal opportunity, and so on. The choice of fairness definition for a deployment context is therefore itself an ethical choice, requiring explicit deliberation by the governance committee responsible for model approval.

The governance architecture proposed in Chapter 6 takes a pragmatic approach to algorithmic fairness, requiring multiple fairness metrics to be computed across population subgroups before deployment authorisation, and requiring human-in-the-loop review for any patient flagged primarily on the basis of missing data. This approach acknowledges that algorithmic fairness cannot be reduced to a single metric, and that the deployment workflow is the appropriate site for fairness safeguards rather than the model architecture alone.

2.6.4 Distributive Justice and AI Adoption

A critical and often neglected dimension of AI governance in public health is the distributional question: who bears the costs of AI investment, and who captures the benefits? In the Zambian HIV programme context, the costs of AI implementation fall primarily on the government, development partners and research institutions, while the benefits accrue primarily to patients, particularly those at risk of treatment interruption, virological failure or TB co-infection. Evidence from this thesis indicates that the patients most likely to benefit from AI-driven identification and support are disproportionately drawn from socioeconomically marginalised groups. AI adoption

designed to address these equity dimensions can therefore serve as a mechanism for reducing health inequalities aligned with SDG 10 [12], [21].

The distributive-justice perspective has practical consequences for the design of the proposed governance architecture. First, it motivates the explicit inclusion of civil society and patient representation in the AI Governance Committee. Second, it motivates equity-disaggregated performance monitoring as a standing reporting requirement. Third, it motivates the inclusion of equity-focused gate criteria in the five-phase adoption roadmap: subgroup performance within defined tolerance is a precondition for progression to the next adoption phase.

2.6.5 Data Sovereignty

A further governance dimension specific to LMIC contexts is data sovereignty: the principle that data generated within a national health system should be governed by national institutions and used in ways aligned with national priorities [54]. Data sovereignty is particularly consequential where international development partners fund the data infrastructure and require access to the data for global programmatic reporting. The CDSS integration architecture proposed in Chapter 6 preserves data sovereignty by operating within Zambia's existing OpenHIE-aligned infrastructure, with no requirement for patient-level data to be transferred outside national institutional control.

2.7 Critical Review and Comparison of Related Works

This section provides the single critical synthesis of the related literature. It combines a structured comparison, the study-by-study mapping in Table 2.3, with a thematic comparison across the principal study domains (Section 2.8), so that the same body of work is examined first for what each study individually contributes and omits, and then for the patterns that recur across domains. The two views are complementary rather than separate reviews, and they converge on the four gaps stated at the end of this synthesis and carried forward in Section 2.9.

Table 2.3 synthesises fifteen of the most relevant prior studies, explicitly mapping what each study contributes and what it does not address. This enables direct identification of the methodological, contextual and ethical gaps that motivate the three empirical studies in this thesis. The table is organised by domain coverage (ML prediction, EHR data, LMIC focus, explainability and ethics) and identifies the principal gap that each study leaves open.

Table 2.3: *Key Studies on AI/ML for EHR-Based Predictive Analytics in Public Health (✓= addressed; ✗ = not addressed; HIC = high-income country).*

Authors	Year	ML Prediction	EHR Data	LMIC	XAI/Ethics	Gap Identified
Rajkomar et al.	2019	✓ Deep learning	✓ Large-scale	✗ HIC only	✗ Limited	Limited LMIC applicability
Beam & Kohane	2018	✓ Conceptual	✓ EHR-centric	✗ No LMIC	✗ Not explored	No empirical LMIC validation
Shickel et al.	2021	✓ Review	✓ Time-series	✗ Limited LMIC	✗ Marginal	Algorithms over deployment
Christodoulou et al.	2019	✓ Comparative	✓ Clinical	✗ No LMIC	✓ Transparency	No public health context
Hasson et al.	2020	✗ No ML	✓ Health IS	✓ LMIC-focused	✗ Limited	No predictive analytics
Mutai et al.	2020	✓ Risk prediction	✓ Routine	✓ LMIC	✗ XAI absent	Limited external validation
Musukwa et al.	2021	✓ ML prediction	✓ National EHR	✓ LMIC	✗ Not discussed	No external validation
Goldstein et al.	2018	✓ Predictive	✓ EHR data	✗ No LMIC	✓ Calibration	Limited operational guidance
Van Calster et al.	2019	✓ Risk theory	✗ Not EHR-specific	✗ No LMIC	✓ Decision reliability	Lacks applied cases
Lundberg & Lee	2017	✓ Explainability	✗ General ML	✗ No LMIC	✓ Strong XAI	No health evaluation
Vayena et al.	2018	✗ No prediction	✗ Conceptual	✓ Global relevance	✓ Ethics-centred	No technical work
WHO	2021	✗ No modelling	✗ Conceptual	✓ LMIC-relevant	✓ Governance	No empirical assessment

Authors	Year	ML Prediction	EHR Data	LMIC	XAI/Ethics	Gap Identified
Rasmy et al.	2023	✓ Deep learning	✓ Large EHRs	✗ High-resource	✗ Limited XAI	High compute demands
Rieke et al.	2020	✓ Privacy-preserving	✓ Multi-site	✓ Potential LMIC	✓ Data sharing	Limited LMIC deployment
Ngcobo et al.	2025	✓ AI for HIV	✓ Review	✓ Global LMIC	✓ Governance	No empirical framework

The table makes visible a structural pattern that motivates the thesis. Studies that contribute methodologically (Rajkomar, Beam & Kohane, Shickel, Christodoulou, Goldstein, Van Calster, Lundberg & Lee, Rasmy) are typically conducted in high-income contexts and either do not address ethics and explainability or do so abstractly. Studies that contribute to LMIC contexts (Hasson, Mutai, Musukwa, WHO, Ngcobo) typically do not address the full methodological depth required for trustworthy ML, and either omit external validation or omit explainability. Studies that contribute to ethics and governance (Vayena, WHO) typically do not contribute empirical evidence. No single study in the table addresses all five dimensions simultaneously. The integrated four-component thesis design described in Chapter 1 is the response to this structural gap.

2.8 Thematic Comparison Across Study Domains

Continuing the single critical synthesis begun in Section 2.7, this section reads the same literature thematically rather than study by study, grouping it under the five domains in which prior work clusters. Where Table 2.3 records what each study does and does not do, the comparison below identifies the patterns that recur across studies within each domain; together they avoid duplicating the review and instead present one analysis from two complementary angles.

2.8.1 Clinical and Population Risk Prediction

The most established body of prior work concerns risk prediction from EHR data, covering mortality, disease progression and treatment outcomes. Large-scale studies from high-income settings demonstrate the feasibility of applying deep learning to longitudinal EHRs [56]–[58], while comparative analyses consistently find that performance gains over simpler models are often

modest [43], [59], [60]. Studies specifically targeting LMIC settings, such as Mutai et al. [49] in Kenya and Musukwa et al. [50] in Zambia, demonstrate feasibility of ML prediction from routine clinical data but highlight persistent gaps in external validation, calibration assessment and explainability evaluation. This thesis directly addresses these gaps through the facility-holdout design and SHAP interpretability framework of Study Two.

2.8.2 Disease Surveillance and Epidemiological Forecasting

A second research stream applies ML to disease surveillance and epidemiological forecasting, often combining EHRs with laboratory reports or syndromic surveillance data [61]–[63]. While such applications show strong short-term predictive performance, concerns persist about model stability, interpretability and transferability across settings [62], [66]. The gap between modelling performance and operational deployment is a recurring theme, with few studies linking forecasting outputs to explicit decision thresholds or public health protocols. The thesis is not directly aimed at the surveillance and forecasting use case, but the methodological contributions on calibration, external validation and threshold optimisation are portable to it.

2.8.3 Data Quality, Preprocessing and Feature Engineering Studies

Data quality challenges are a pervasive theme in the prior literature, particularly for LMIC contexts. The Weiskopf and Weng framework [16] remains the most-cited reference for EHR data quality assessment in clinical research, but as Section 2.4 noted, it does not address several dimensions essential for AI training. Dong et al. [68] proposed a framework specifically for predictive analytics that includes some of the dimensions added in this thesis, but does not include the missingness-outcome confounding test that has proved diagnostically central to the empirical findings. Wang et al. [70] systematically demonstrated the sensitivity of ML predictions to imputation choice in EHR studies, underscoring the importance of explicit preprocessing pipelines. This thesis addresses these gaps directly through the data quality assessment in Study Three and the leakage-safe temporal feature engineering in Study Two.

2.8.4 Model Evaluation and Validation Practices

Evaluation practices in the AI/ML health literature have been widely critiqued for over-reliance on discrimination metrics, particularly AUC-ROC, at the expense of calibration and external

validation [41], [43], [70]. The thesis contributes a facility-holdout external validation design that provides a conservative and realistic assessment of deployability, and reports calibration explicitly. The systematic review (Chapter 5) confirms that this combination of features is rare in the published literature on LMIC HIV care prediction.

2.8.5 Explainability, Ethics, Governance and 4IR Policy Studies

The prior literature on explainability, ethics, governance and 4IR policy has largely developed in parallel to the methodological literature, with limited integration between the two streams. Lundberg and Lee's SHAP paper [52] is the methodological foundation for the population-level and patient-level explanations in this thesis. Vayena et al. [53] and WHO [54] provide the ethical and governance frameworks that motivate the equity safeguards in the proposed deployment architecture. Ngcobo et al. [55] provide a recent and directly comparable systematic review of AI for HIV care in sub-Saharan Africa, identifying many of the same gaps that motivate this thesis but proposing no empirical framework. The thesis can be read as an empirical and architectural response to the Ngcobo et al. agenda.

Drawing the study-by-study mapping of Table 2.3 together with this thematic comparison, the critical review converges on four gaps that no single prior study closes and that, jointly, define the space this thesis occupies. First, an evaluation gap: across the prediction and validation literature, models are appraised predominantly on discrimination, while external validation, calibration and explainability, the properties that determine whether a model is safe to act on, are reported only sporadically. Second, a context gap: the great majority of methodologically strong work originates in high-income settings with dense data, leaving LMIC routine-EHR settings, where data are sparse and confounded, substantially underrepresented and the portability of published methods untested. Third, an explainability-and-governance gap: the interpretability and the ethics/governance literatures have developed in parallel rather than in combination, so that few studies link patient-level explanation to the governance mechanisms (fairness gating, human-in-the-loop review) needed to use it responsibly. Fourth, a deployment gap: almost no study specifies the ICT architecture and institutional pathway required to move a research-grade model into routine clinical use under real infrastructure constraints. These four gaps are not independent shortcomings of separate studies but a single, connected deficiency in the field, and it is their

conjunction that motivates the integrated, four-component response set out in the remainder of this chapter and formalised in Section 2.9.

2.9 Identified Gaps in the Literature

Across the six thematic domains reviewed and the fifteen studies tabulated in Section 2.7, four persistent gaps motivate the empirical studies and the extended governance analysis in this thesis. These gaps were introduced in Chapter 1 as the problem statement and are restated here, after the empirical literature review, with reference to the specific studies that exemplify each gap.

Gap 1: Evaluation imbalance. Discrimination metrics predominate at the expense of calibration and external validation, limiting the operational credibility of published models [41], [43], [70]. The systematic review (Chapter 5) confirms that calibration is reported in only 6.2% of studies and external validation in only 7.8%.

Gap 2: LMIC underrepresentation. Only 18.8% of recent studies were conducted in LMIC settings, despite these settings carrying the greatest burden of preventable disease [24], [28]. Even among studies set in LMICs, explainability assessment is reported in only 8.3% of cases compared with 36.4% in high-income studies.

Gap 3: Explainability and governance gap. Explainability and governance frameworks are treated as conceptual addenda rather than integral evaluation criteria [53], [54], [74]. No single study in the comparative table (Section 2.7) addresses all five dimensions of methodology, EHR data, LMIC focus, explainability and ethics simultaneously.

Gap 4: Deployment gap. The ICT architecture required to operationalise ML frameworks in routine clinical settings has not been systematically addressed for the Zambian context, creating a gap between research-grade pipelines and production-ready CDSS that the proposed five-layer integration architecture (Chapter 6) is designed to close.

2.10 Conceptual and Theoretical Framework

As introduced in Section 1.7, the conceptual framework for this thesis integrates three theoretical perspectives: post-positivism in health informatics research, the socio-technical systems perspective on health information technology, and the equity-and-distributive-justice perspective on AI in public health [47], [53], [75]. Rather than restating those perspectives here, this section develops them further by supplying the named information-systems theory that grounds the governance and deployment components of the thesis.

The socio-technical perspective is given specific theoretical content by drawing on established information-systems (IS) theory, which is necessary to ground the governance and deployment components that Chapters 5 and 6 advance. Three traditions are pertinent. The DeLone and McLean IS Success Model frames the success of an information system as a multidimensional construct in which information quality, system quality and service quality jointly determine use, user satisfaction and ultimately net benefits; it is directly applicable here because it predicts that a technically accurate model embedded in a low-information-quality data environment, of the kind the data-quality assessment documents, will not produce net clinical benefit unless the data-quality dimension is remediated, which is precisely the rationale for treating the eight-dimension DQA as a deployment gate rather than a preliminary check. The Technology Acceptance Model (TAM) and its later synthesis, the Unified Theory of Acceptance and Use of Technology (UTAUT), explain the behavioural determinants of whether clinicians will actually adopt a decision-support tool, identifying perceived usefulness and ease of use (TAM) and, in UTAUT, performance expectancy, effort expectancy, social influence and facilitating conditions as the proximal drivers of acceptance and use. These constructs motivate design choices in the proposed governance and deployment architecture that would otherwise appear arbitrary: explainability provisions and human-in-the-loop review address performance expectancy and trust; workflow integration and CPU-only, offline-resilient operation address effort expectancy and facilitating conditions; and the stakeholder-engagement and clinical-champion provisions address social influence.

Taken together, these IS theories supply the named theoretical tradition that grounds the thesis's "4IR health transformation framework" and "trustworthy AI framework": the DeLone and McLean model anchors the claim that data and system quality are preconditions for benefit, while TAM/UTAUT anchor the claim that governance, explainability and deployment design are

determinants of real-world use rather than optional additions. The governance architecture in Chapter 6 can therefore be read as an applied IS-success and technology-acceptance intervention specialised to the constraints of a national HIV programme, rather than as a free-standing set of recommendations.

Operationalising these three perspectives requires the integrated four-component design described in Chapter 1: the systematic review provides the methodological landscape (post-positivism); the empirical ML modelling demonstrates the technical feasibility (socio-technical systems); the data quality assessment provides the empirical foundation for equity safeguards (equity-and-distributive-justice); and the governance and ICT architectures translate the technical work into operational practice (socio-technical and equity-and-distributive-justice combined). The systematic review findings are presented as Study One in Chapter 5; the PRISMA 2020 study selection flow diagram and the annual distribution of included studies are introduced in Sections 5.2.1 and 5.2.2.

The conceptual framework also motivates the integration of empirical and stakeholder-engagement evidence. The two dissemination meetings convened during the research — with the Ministry of Health and Provincial Health Office M&E teams; with the CIDRZ AI Think-Tank — provided structured qualitative input that has been integrated into the recommendations of Chapter 7 and is reported in detail in Appendix F. The conceptual commitment to socio-technical systems thinking requires that this stakeholder input be treated as evidence on equal footing with the empirical data, not as commentary on the empirical findings.

2.11 Proposed Framework Components

Drawing on the gaps identified in Section 2.9 and the conceptual framework articulated in Section 2.10, the thesis proposes a four-component integrated framework that is operationalised across the empirical and integrative chapters.

2.11.1 Component One: Methodological

The methodological component centres on a leakage-safe, multi-outcome, externally validated and explainable ML framework for HIV outcome prediction. The component is operationalised in Chapter 3 (Section 3.4) as the temporal feature engineering pipeline, the stacked ensemble

architecture and the F1-optimised threshold selection; its empirical results are reported in Chapter 5 (Sections 5.3 and 5.4). The methodological component addresses Gap 1 (evaluation imbalance) and Gap 3 (explainability gap) directly.

2.11.2 Component Two: Data Quality

The data quality component centres on an eight-dimension DQA framework with missingness-confounding analysis. The component is operationalised in Chapter 3 (Section 3.5) as the eight-dimension DQA framework definition, the plausibility range specifications and the missingness-outcome confounding test; its empirical results are reported in Chapter 5 (Section 5.5). The component addresses Gap 2 (LMIC data underrepresentation) by establishing a reproducible diagnostic toolkit that can be applied across LMIC HIV programmes, and addresses Gap 3 (governance gap) by providing the empirical foundation for the fairness safeguards in the proposed governance architecture.

2.11.3 Component Three: Governance

The governance component centres on an institutional architecture for AI governance and an eight-requirement minimum standard for data quality. The component is operationalised in Chapter 6 (Sections 6.4 and 6.5) and synthesised into the recommendations of Chapter 7 (Section 7.3). The governance component addresses Gap 3 directly and provides the institutional foundation for sustainable AI adoption.

2.11.4 Component Four: Deployment

The deployment component centres on a five-layer ICT architecture for CDSS integration with offline-resilient design choices appropriate to Zambia's connectivity reality. The component is operationalised in Chapter 6 (Section 6.5) and detailed in the companion ICTAZ journal publication [82]. The deployment component addresses Gap 4 directly and provides the engineering pathway from research to operational AI.

Each component addresses one of the four gaps identified in Section 2.9, and together they constitute a coherent pathway from research to responsible deployment. The four components are

presented in detail across Chapters 3 to 5, with the cross-component synthesis matrix reported in Section 5.6 of Chapter 5.

2.12 HIV Epidemiology, Care Cascade and the Zambian Context

2.12.1 Global and Sub-Saharan African Burden

UNAIDS estimates indicate that 39.0 million people were living with HIV at the end of 2023, with sub-Saharan Africa accounting for 25.6 million (approximately 65.6%) of the global total [8]. Within sub-Saharan Africa, eastern and southern Africa account for the highest absolute burden, with Zambia, Malawi, Zimbabwe, Mozambique, South Africa, Botswana, Lesotho, Eswatini, Tanzania, Uganda, Kenya and Rwanda contributing the bulk of the regional total. New HIV infections globally fell from 2.5 million in 2010 to approximately 1.3 million in 2023, a 48% reduction over the period, but the rate of decline has slowed since 2020, with new infections plateauing in eastern and southern Africa while declining more substantially in western and central Africa [8].

The epidemiological transition within sub-Saharan Africa's HIV epidemic has consequences for AI deployment priorities. The initial decades of the epidemic were dominated by efforts to scale antiretroviral therapy coverage, where the operational priority was reaching patients newly diagnosed and getting them onto treatment. As coverage has reached or approached the UNAIDS 95-95-95 targets in several countries, the operational priority has shifted to retention in care, viral suppression and prevention of new infections among key populations. This shift creates demand for predictive analytics targeting the operationally consequential rare events (treatment interruption, virological failure, key population non-engagement) that this thesis addresses, rather than the high-prevalence events (HIV-positive status, ART initiation) that dominated earlier intervention generations.

2.12.2 The HIV Care Cascade and Operational Decision Points

The HIV care cascade, developed by Gardner et al. and operationalised globally through the UNAIDS 95-95-95 framework, identifies five operational decision points where AI-driven prediction can add value: HIV testing (where AI can target testing to high-risk individuals); ART initiation (where AI can predict same-day initiation feasibility); retention in care (where AI can

predict interruption risk); viral suppression (where AI can predict virological failure); and onward transmission risk (where AI can support prevention prioritisation). The empirical work of this thesis addresses three of these five decision points (retention via IIT12m, viral suppression via UVL12m, and one of the principal differential diagnoses for adverse outcomes via TB12m).

The cascade framework also identifies the disproportionate impact of late-stage cascade failures: a patient who interrupts treatment at month 18 of ART represents a much larger loss to programme effectiveness than a patient who never starts ART, because the treatment interruption is preceded by 18 months of investment in counselling, drug supply and clinical monitoring. Predictive identification of impending interruption therefore captures more programmatic value per intervention than predictive identification of test-positive patients at the same risk level. This is a foundational economic argument for the prioritisation of retention-focused AI deployment, and is the basis for Recommendation R9 in Chapter 7, which prioritises IIT12m as the operational entry point for clinical AI deployment.

2.12.3 HIV-TB Co-Infection and the Tuberculosis Outcome Rationale

Tuberculosis remains the leading cause of HIV-related mortality globally, accounting for approximately one in three deaths among PLHIV. The reciprocal interaction between the two infections is well established: HIV-induced immune suppression increases the risk of progression from latent to active tuberculosis, while active tuberculosis accelerates HIV disease progression and increases mortality risk independently of CD4 count. PLHIV face a 20- to 30-fold higher risk of active TB than the HIV-negative population, a multiplier that has not changed substantially with the scale-up of antiretroviral therapy because immune reconstitution is incomplete in many patients and because TB infection rates remain elevated in high-burden settings even with broad ART coverage.

The TB12m outcome operationalised in this thesis (incident TB or TB treatment initiation within 12 months of the patient's index date) is intended to support targeted intensified case-finding and isoniazid preventive therapy (IPT) eligibility decisions. WHO guidelines recommend symptom-based screening at every clinical encounter for PLHIV, with IPT offered to all those who screen negative for active TB, but in practice IPT coverage and active case-finding completeness are below the levels assumed in modelling exercises. An AI-driven risk score that prioritises high-risk patients for more intensive screening, including molecular diagnostics where available, can lift the

overall programmatic yield without proportionally increasing the resource burden. The TB12m AUC-ROC of 0.707 achieved in this thesis under facility-holdout external validation is within the operational range that supports this kind of targeted screening.

The label conflation issue noted in Section 2.3.4 is, however, a substantive limitation of the current operationalisation. Of the 36% overall TB prevalence in the cohort, an unknown fraction reflects prevalent TB (already present at the index date) rather than incident TB (newly diagnosed within the 12-month window). This conflation reflects the structure of available SmartCare data fields rather than a modelling choice. The Recommendation R6 in Chapter 7 (standardised schema across SmartCare facilities) is partially aimed at resolving this issue through more granular outcome definitions in future operationalisation.

2.12.4 Treatment Interruption and Retention

Treatment interruption is operationally defined in Zambia's national HIV programme as a clinic visit gap of more than 90 days (in some definitions, more than 60 days) without documented transfer or self-discharge. The interrupted patient is generally counted as "lost to follow-up" (LTFU) for programme reporting purposes, and is then either returned to care (with re-engagement support) or formally exited as "lost". The IIT12m outcome operationalised in this thesis (any classification of "inactive" within 12 months of the index date) captures both interrupted patients who are subsequently returned to care and those who are exited as lost, providing a unified operational target for retention support.

Predictors of interruption documented in the prior literature include demographic factors (younger age, male sex, lower educational attainment), clinical factors (longer time since enrolment, higher viral load at index, lower CD4 count), engagement factors (history of missed visits, late presentation for prior appointments) and contextual factors (distance to facility, recent relocation, key population status). The leakage-safe feature engineering of this thesis captures the demographic, clinical and engagement factors directly from EHR data; contextual factors are only partially captured (through facility identifiers and household income where recorded). Extension to include explicit contextual variables (e.g., GPS-based distance to facility, mobile phone access) is a natural future research direction and is included in Future direction 5 of Chapter 7.

2.12.5 Virological Failure and the UVL Outcome

Viral load suppression is the primary biological marker of ART effectiveness. Patients with sustained viral suppression (typically defined as <1,000 copies/mL, or <200 copies/mL under more stringent definitions) are at very low risk of HIV-related morbidity, mortality and onward transmission. The UNAIDS 95-95-95 target of 95% of patients on ART being virally suppressed is the most operationally consequential of the three targets, because suppression is the proximate endpoint that delivers public health value.

Virological failure in PLHIV on ART arises from several mechanisms: poor adherence (the dominant cause); drug resistance (developing over time, particularly in patients with suboptimal adherence or interrupted treatment); drug-drug interactions; and pharmacokinetic variability. Predictive identification of patients at risk of virological failure supports targeted enhanced adherence counselling, drug resistance testing where available, and regimen optimisation. The UVL12m outcome operationalised in this thesis captures any viral load ≥1,000 copies/mL within 12 months of the index date, which is a programmatically meaningful definition of virological failure for first-line ART monitoring.

The empirical UVL12m findings of this thesis (XGBoost AUC-ROC 0.781, AUC-PR 0.064) represent meaningful predictive performance for an outcome that prior LMIC studies have generally not predicted with the equivalent rigour. The dominance of education in the SHAP and permutation importance analyses, combined with the missingness-outcome confounding finding, requires interpretation through the equity lens developed in Chapter 6.

2.12.6 Differentiated Service Delivery and the Operational Context

Differentiated service delivery (DSD) is the dominant operational paradigm for HIV care in sub-Saharan Africa since approximately 2015 [19], [87]. The principle of DSD is to match the intensity of services to patient risk: stable, well-engaged patients receive less intensive care (e.g., multi-month dispensing, fast-track refill, community pickup), while patients at higher risk or with greater clinical complexity receive more intensive care (e.g., monthly clinic visits, enhanced adherence counselling, peer support). DSD has been shown to reduce facility congestion, improve patient satisfaction and maintain or improve clinical outcomes across multiple sub-Saharan African settings [19], [87].

The integration of AI-driven risk prediction into DSD is a natural fit: the risk score generated by the predictive model can become an explicit input to the DSD assignment decision, supplementing the clinical judgement currently applied by clinicians. This integration was specifically discussed at the first dissemination meeting (Appendix F.1) and is the operational rationale for Recommendation R11 in Chapter 7. The joint workshop with the DSD Technical Working Group agreed at that meeting will operationalise the integration during Phase 2 of the adoption roadmap.

2.14 AI Ethics, Governance and the Trustworthy AI Framework

2.14.1 The Trustworthy AI Movement

The trustworthy AI movement has emerged over the past decade as the dominant normative framework for AI deployment in high-stakes domains, including health, finance, criminal justice and education. The movement integrates ethical principles (beneficence, non-maleficence, autonomy, justice, explicability) with technical requirements (reliability, safety, fairness, privacy, transparency) and governance mechanisms (accountability, auditability, regulatory oversight). Influential frameworks include the European Union's Ethics Guidelines for Trustworthy AI (2019), the OECD's AI Principles (2019), the WHO's Ethics and Governance of AI for Health (2021) and the EU AI Act (2024). For health specifically, the most directly relevant frameworks are WHO's guidance on AI ethics and governance [54] and the recent regulatory considerations on AI for health [107].

To avoid any ambiguity about the role this material plays, this section is positioned as part of the literature review: it surveys the external trustworthy-AI and AI-ethics landscape against which the thesis is situated, and it is not itself the thesis's theoretical framework. The operative theoretical framework, the post-positivist, socio-technical (IS-success and technology-acceptance) and equity-and-distributive-justice lens through which the empirical work is designed and interpreted, is the one set out in Section 2.10; the trustworthy-AI principles reviewed here function as an external normative benchmark that the framework is later shown to satisfy, rather than as the lens itself. Read in that way, the present section informs the governance requirements without competing with Section 2.10 for the role of theoretical foundation.

The trustworthy AI framework converges on six core principles: (i) human agency and oversight; (ii) technical robustness and safety; (iii) privacy and data governance; (iv) transparency and

explainability; (v) diversity, non-discrimination and fairness; and (vi) societal and environmental well-being. Each of these principles maps to specific design decisions in the AI deployment lifecycle, from data sourcing through model development to deployment and monitoring. The thesis framework addresses all six principles in the four-component architecture: methodological (transparency, explainability, robustness), data quality (privacy, fairness foundations), governance (oversight, accountability, diversity) and deployment (safety, well-being).

2.14.2 Algorithmic Fairness: Mathematical and Operational Definitions

Algorithmic fairness is a particularly active area of research with multiple mathematical definitions that are mutually incompatible in general. The principal definitions used in the health AI literature are: (i) demographic parity, requiring that the proportion of positive predictions is equal across subgroups; (ii) equalised odds, requiring that both the true positive rate and false positive rate are equal across subgroups; (iii) equal opportunity, requiring only that the true positive rate is equal across subgroups; and (iv) calibration parity, requiring that calibration is equal across subgroups [45], [46], [126]. Each definition reflects a different ethical commitment, and the choice between them is a value judgement rather than a technical one.

The thesis takes a pragmatic approach to algorithmic fairness, requiring multiple fairness metrics to be computed and reported across population subgroups before deployment authorisation, and reserving the choice of operational threshold for the governance committee. This approach acknowledges that algorithmic fairness cannot be reduced to a single metric, and that the deployment workflow is the appropriate site for fairness safeguards rather than the model architecture alone. The companion equity-disaggregated performance monitoring framework [139] provides the methodological foundation for the equity monitoring required by the governance architecture.

2.14.3 Bias Sources in Health AI

Sources of bias in health AI deployment are now well documented [46], [47], [74], [126]. They include: (i) selection bias in training data (e.g., patients with worse documentation are systematically under- or over-represented); (ii) measurement bias in features and outcomes (e.g., different facilities use different diagnostic criteria); (iii) historical bias reflecting structural inequalities in healthcare access (e.g., the Obermeyer et al. [47] finding that historical spending

underestimates health needs for Black patients); (iv) deployment bias (e.g., the model is used differently across subgroups, with different consequences); and (v) feedback bias (e.g., the model's outputs influence future data, creating a self-reinforcing pattern).

The empirical findings of this thesis surface several of these bias mechanisms directly. The feature coverage collapse (Chapter 5, Section 5.5.3) reflects selection bias in the data infrastructure: 87.8% of patients have only demographic features available, and the patients with denser feature coverage are systematically those who present more frequently for care, which is itself correlated with socioeconomic status, geographic accessibility and clinical complexity. The missingness-outcome confounding (Chapter 5, Section 5.5.7) reflects measurement bias at the facility level: facilities with more thorough data collection also have different outcome rates, generating spurious correlations that the model may learn as if they were genuine clinical signal. The dominance of education in UVL12m predictions (Chapter 5, Section 5.4.2) is consistent with historical bias: educational attainment correlates with structural factors that affect both data completeness and clinical outcomes, in ways that are difficult to disentangle in observational data.

2.14.4 Privacy, Data Sovereignty and Consent

Privacy and data sovereignty are particularly consequential for LMIC AI deployment because the data infrastructure is often funded and partially controlled by international development partners, while the data themselves are generated within national health systems and pertain to national citizens. The principle of data sovereignty asserts that data generated within a national health system should be governed by national institutions and used in ways aligned with national priorities [54]. This principle has practical implications for the proposed CDSS integration architecture: patient-level data should not be transferred outside national institutional control without explicit national authorisation; model training should be conducted on national infrastructure or with strict data residency guarantees; and the AI Governance Committee should be composed primarily of national institutional representatives.

Patient consent for AI-based decision support presents particular complexities. The standard HIV care consent at enrolment in Zambia's SmartCare programme covers the use of routine data for clinical care and public health learning, but does not explicitly mention AI-based prediction or risk scoring. The proposed CDSS architecture (Chapter 6) preserves this consent framework by treating AI-based risk scores as an extension of clinical decision support, with the same governance and

oversight as other clinical decision support tools. However, transparency to patients about the use of AI is itself an ethical requirement: patients have the right to know when their care is being informed by algorithmic prediction, and to opt out where they choose. The Ploug and Holm [124] analysis of the right to refuse AI-based diagnostics and treatment planning provides the philosophical foundation for this transparency requirement. Operationalising patient transparency without overwhelming patients or undermining care delivery is a specific subject for Phase 2 of the deployment roadmap.

2.14.5 Regulatory Landscape for AI in Health

The regulatory landscape for AI in health is rapidly evolving. In the European Union, the AI Act (Regulation EU 2024/1689) [108] classifies AI systems by risk category, with high-risk systems including those used for medical diagnosis or treatment decision support subject to mandatory pre-market conformity assessment, post-market monitoring and registration in an EU database. In the United States, the FDA's AI/ML Action Plan [109] outlines a regulatory pathway for AI/ML-based software as a medical device (SaMD), with provisions for adaptive learning algorithms that change over time. The WHO's regulatory considerations on AI for health [107] provide guidance for low- and middle-income country regulators developing national regulatory frameworks.

In Zambia specifically, the regulatory framework for clinical AI is still emerging. The Health Professions Council of Zambia and the Pharmacy and Poisons Board (now part of the Zambia Medicines Regulatory Authority) regulate medical devices and clinical decision support tools, but as of 2026 there is no specific AI-focused regulatory pathway. The proposed AI Governance Committee (Chapter 6, Section 6.4.3) is intended in part to provide the institutional foundation for AI regulation to develop, with the Committee acting as a de facto regulator pending the development of formal regulatory pathways. The Recommendation R1 in Chapter 7 calls explicitly for the establishment of this Committee within twelve months of thesis approval.

2.15 Chapter Summary

This chapter has reviewed the prior literature on ML and AI for EHR-based public health prediction across six thematic domains. The central finding is a structural imbalance: methodological investment in algorithmic development has substantially outpaced investment in evaluation rigour, explainability, governance and deployment architecture, and LMICs, which bear

the greatest burden of preventable disease, are precisely the settings where this imbalance is most acute and where the 4IR opportunity is most consequential.

Four cross-cutting gaps emerge: evaluation imbalance, LMIC underrepresentation, the explainability and governance gap, and the deployment gap. Each gap is mapped to one of the four components of the proposed framework: methodological, data quality, governance and deployment. The conceptual framework integrates post-positivism, socio-technical systems thinking and the equity-and-distributive-justice perspective, motivating the integrated empirical and stakeholder-engagement design of the thesis. The next chapter presents the research methodology for all empirical studies, grounded in the gaps and framework established here, and introduces the proposed ICT integration architecture for clinical deployment.

The decisive conclusion of this review is not merely that these four gaps exist, but that they are interdependent and cannot be closed in isolation: an evaluation-rigorous model trained on uncertified data inherits that data's unreliability; data-quality certification without explainability cannot reveal when predictions are driven by missingness rather than clinical signal; and explainable, well-evaluated models without a governance and deployment architecture never reach the point of care. Taken together, therefore, the gaps justify a thesis structured as a single integrated response rather than four separate studies: a methodologically rigorous, leakage-safe and externally validated modelling component (addressing the evaluation gap), built on an eight-dimension data-quality certification component (addressing the data-quality gap), interpreted through an explainability component linked to that certification (addressing the explainability gap), and operationalised through a governance and ICT deployment architecture grounded in the IS-success and technology-acceptance theory set out in Section 2.10 (addressing the deployment and governance gap). This is the structure of the empirical and integrative chapters that follow, and it is the sense in which the four gaps, jointly, determine the design of the thesis.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Research Design and Approach

3.1.1 Overall Design

This thesis adopts a mixed-methods research design that integrates a systematic review (Study One), an applied quantitative modelling study (Study Two) and a systematic data quality assessment (Study Three). The studies are complemented by an economic analysis, governance framework development and ICT architecture design drawing on evidence from the companion journal publications. The principal epistemological stance is post-positivist with a pragmatist orientation: the research assumes that objective knowledge about health system phenomena can be obtained through systematic, empirically grounded inquiry, while acknowledging that measurement and inference are subject to systematic constraints and uncertainties [75], [76]. The pragmatist orientation admits both quantitative and qualitative evidence on the basis of fit-for-purpose rather than commitment to a single ontological or epistemological stance.

The studies are sequentially linked: the systematic review (Study One) establishes the methodological landscape and identifies the gaps that Study Two and Study Three then address; Study Two (modelling) develops the predictive analytic artefact whose facility-level performance is then explained, in part, by the facility-level data quality scores generated in Study Three. The studies share the same six-facility study population, which provides analytic coherence and enables cross-study triangulation.

Figure 3.1 illustrates the research design and methodological approach, showing the relationship between the empirical studies, the 4IR governance and economic analysis, and the ICT architecture, and their integration into the overall thesis contribution.

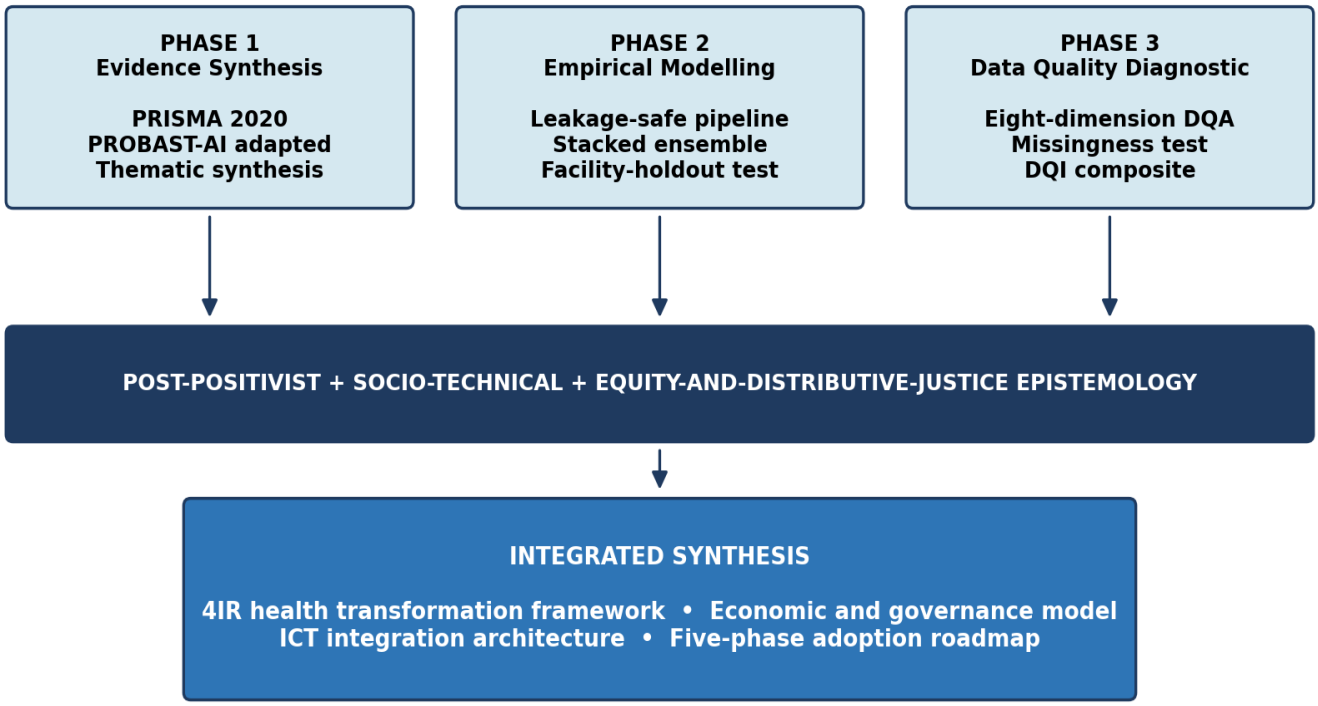

Figure 3.1: *Research design and methodological approach showing the integration of all inter-related empirical studies, the 4IR governance analysis and the ICT deployment architecture into the overall thesis contribution.*

3.1.2 Philosophical Position

The post-positivist position is appropriate because the central empirical questions of the thesis — how well does the ML framework discriminate, how complete is the EHR data, what features dominate predictions — admit quantitative answers that are stable across reasonable variations in analytical choices. The pragmatist orientation is appropriate because the central deployment questions — what governance architecture will work, what funding pathway is feasible, what stakeholder concerns must be addressed — require integration of quantitative empirical findings with qualitative stakeholder evidence and contextual judgement that no single epistemology can supply on its own.

The choice of pragmatism warrants explicit defence against the most credible alternative for mixed-methods health-informatics research, namely critical realism. Critical realism is attractive in this domain because it posits a stratified ontology in which real but unobservable generative mechanisms (here, the social and structural processes that produce both disease risk and the data that record it) underlie observable events, and it would offer a principled account of why, for

example, an education-missingness indicator can stand in for deprivation in the prediction of viral non-suppression. That explanatory ambition is genuine, and this thesis does not dispute the realist claim that such mechanisms exist. The objection is methodological rather than ontological: a critical-realist programme directs the inquiry toward retroductive identification and causal explanation of those mechanisms, whereas the questions this thesis must answer are predictive and operational — whether a leakage-safe model generalises across facilities, whether routine data are fit for training, and whether a governance and deployment architecture is institutionally feasible. None of these is, in the first instance, a claim about underlying causal mechanisms, and treating them as such would subordinate the practical deployment question to an explanatory programme that the available routine data cannot adequately support.

Pragmatism is therefore adopted not as a default but because it takes the research question, rather than a prior ontological commitment, as the arbiter of method, and because the thesis pursues a consequential aim: a trustworthy, deployable decision-support framework. On a pragmatist account, the value of a feature, a model or a governance provision is judged by its consequences for that aim under real-world constraints, which licenses the deliberate mixing of quantitative modelling, quantitative data-quality assessment and qualitative stakeholder evidence within one design. Critically, adopting pragmatism does not discard the realist insight: the missingness-outcome confounding analysis (Section 3.5.4) is precisely the point at which the design surfaces a candidate generative mechanism, and the equity safeguards in the deployment architecture are the practical response to it. In this sense the pragmatist frame absorbs what is most useful in critical realism — attentiveness to the structural processes behind the data — while keeping the inquiry anchored to the predictive and deployment questions the thesis is actually positioned to answer.

The equity-and-distributive-justice perspective described in Section 2.10 of Chapter 2 operates as a normative constraint on the design rather than as an epistemological claim. It requires that the analyses be conducted in ways that surface, rather than hide, equity-relevant patterns; that the reporting be conducted in ways that make equity implications visible to stakeholders; and that the recommendations be designed in ways that address rather than entrench inequalities. The eight-dimension DQA framework, the SHAP-based interpretability framework and the human-in-the-loop review provision in the proposed deployment architecture are each operational consequences of this normative commitment.

3.1.3 Integration of Quantitative and Qualitative Evidence

Two formal dissemination meetings convened during the research provided structured qualitative input that has been integrated into the recommendations of Chapter 7 and is reported in detail in Appendix F. The first meeting engaged the Ministry of Health and Provincial Health Office M&E Data Management teams on the empirical findings and the proposed governance roadmap. The second meeting engaged the CIDRZ AI Think-Tank Team on the methodological choices and the model card development. The qualitative analysis approach for these meetings was structured thematic synthesis: rapporteur notes were structured against pre-defined themes (data quality, methodology, deployment, equity, funding), and emerging themes were identified through iterative coding. The synthesised feedback is integrated into Chapter 6 (Section 6.6) and Chapter 7 recommendations.

3.2 Methodological Framework

3.2.1 Mapping of Research Questions to Studies

The research questions introduced in Chapter 1 are mapped to the empirical studies in a structured way. Table 3.1 presents the mapping for Study One; the corresponding mappings for Studies Two and Three are integrated into Sections 3.4 and 3.5 respectively.

Table 3.1: *Mapping of research questions to review processes for Study One.*

Research Question	Review Process
RQ1.1: ML algorithms in LMIC settings	Extraction of ML model types, algorithm categorisation, comparison of reported performance metrics, assessment of interpretability and computational feasibility
RQ1.2: Preprocessing and feature engineering	Extraction of preprocessing methods, feature engineering strategies, handling of missingness, heterogeneity and temporal data
RQ1.3: Evaluation and validation	Extraction of evaluation metrics, validation strategies, calibration reporting, explainability methods and governance considerations

3.2.2 Research Quality Assurance

Three quality-assurance mechanisms operate across the methodological framework. First, all data processing and analytical code was version controlled, with each analytical step traceable to a specific commit and reproducible from a fixed random seed (seed 42). Second, all preprocessing parameters were fitted exclusively on the training set and applied to the test set without refitting, preventing test-set contamination. Third, all reported metrics were computed on the held-out test set comprising the two Matero facilities, with no test-set patient ever contributing to model fitting. These three mechanisms together provide the methodological foundation for the strong external validity claims advanced in Chapter 5.

3.2.3 Documentation and Reproducibility Standards

The methodology follows the TRIPOD reporting guideline for clinical prediction models [71] and anticipates the TRIPOD-AI extension currently in development [72]. The DQA reporting follows an extension of the Weiskopf and Weng framework [16] developed in the companion DQA paper. The systematic review reporting follows PRISMA 2020 [20]. The reproducibility appendix (Appendix E) documents the software environment, library versions, random seeds and code availability arrangements that support independent re-implementation of all empirical studies.

3.3 Study One: Systematic Review Methodology

3.3.1 Protocol and Reporting Standards

The systematic review was conducted in accordance with the PRISMA 2020 framework [20], with adapted PROBAST [21] risk-of-bias assessment extended to ML-specific considerations [22]. A pre-specified protocol was developed prior to the search execution, covering the research questions, search strategy, inclusion and exclusion criteria, screening procedures, data extraction template, risk-of-bias assessment approach and intended synthesis methods. The full protocol is summarised in Appendix B.

3.3.2 Search Strategy and Selection Criteria

A comprehensive literature search was conducted across five major academic databases: IEEE Xplore, Scopus, Web of Science, ACM Digital Library and PubMed. The primary search string combined controlled vocabulary and free-text terms relating to AI, ML, EHRs, predictive analytics, public health, epidemiology and resource-limited settings, restricted to peer-reviewed

English-language publications from January 2018 to March 2025. The representative search string adapted for PubMed is presented in Appendix B.

Study selection followed a three-stage PRISMA-consistent screening process: title and abstract screening, full-text screening and final inclusion. Two reviewers screened each record independently, with disagreements resolved by discussion or, where necessary, by reference to a third reviewer. Table 3.2 summarises the inclusion and exclusion criteria applied.

Table 3.2: *PRISMA-compliant inclusion and exclusion criteria for the systematic review.*

Category	Inclusion Criteria	Exclusion Criteria
Publication year	2018–2025	Pre-2018
Language	English only	Non-English
Study type	Peer-reviewed articles, systematic reviews, validated conference papers	Editorials, opinion pieces, blogs
Focus	AI/ML applied to public health or epidemiology	Purely clinical AI without population focus
Methods	Empirical AI models, epidemiological simulations, decision-support systems	Theoretical AI papers without health application
Setting	LMIC (primary); high-income as comparative baseline	No geographic or health system context reported
Data	Electronic health records or routinely-collected clinical data	Synthetic or laboratory-only datasets

3.3.3 Database-Specific Search Yield

Table 3.3 presents the database-specific search yields, with deduplication reducing the 1,847 initial records to 1,260 unique records for screening.

Table 3.3: *Database-specific search strings and yield.*

Database	Records Identified	After Deduplication
IEEE Xplore	312	291
Scopus	521	358

Database	Records Identified	After Deduplication
Web of Science	428	264
ACM Digital Library	298	186
PubMed	288	161
Total	1,847	1,260

3.3.4 Data Extraction Template

A structured data extraction template captured the following dimensions for each included study: study setting and population; data sources and sample size; prediction task and outcome; ML algorithms used; preprocessing and feature engineering methods; evaluation metrics and validation strategy; explainability, ethical and governance considerations; and reported limitations. The template was piloted on ten studies before full-scale application and refined based on inter-reviewer agreement assessment.

3.3.5 Risk-of-Bias Assessment

Quality and risk-of-bias assessment was conducted using an adapted PROBAST framework extended to ML-specific considerations [21], [22]. Table 3.4 summarises the PROBAST-AI-adapted risk-of-bias domains used in the assessment.

To ensure the reliability of the risk-of-bias judgements, each included study was assessed independently by two reviewers, who rated every PROBAST-AI domain without reference to the other's ratings. Inter-rater agreement before reconciliation was quantified using Cohen's kappa, computed across the domain-level judgements, with agreement interpreted on the conventional scale in which values above 0.6 denote substantial and above 0.8 almost-perfect agreement. Disagreements were resolved through a structured adjudication procedure: the two reviewers first discussed each discordant domain to reach consensus, and any judgement that remained unresolved was referred to a third senior reviewer whose decision was final. This independent dual-assessment and adjudication process, rather than single-reviewer rating, is the standard expected of a PRISMA-compliant review and guards against idiosyncratic or systematically lenient bias appraisal; the reconciled domain ratings are those reported in the synthesis (Section 3.3.6) and tabulated in Appendix B.5.

Table 3.4: *PROBAST-AI adapted risk-of-bias domains.*

Domain	Assessment Focus
Participants	Population definition, inclusion criteria, source data quality
Predictors	Predictor definition, measurement validity, blinding to outcome
Outcome	Outcome definition, measurement validity, blinding to predictors
Sample size	Adequacy of sample size for model development and validation
Missing data	Handling of missing data, imputation choices, sensitivity analyses
Statistical analysis	Model specification, performance metrics, calibration, decision curves
Validation	Internal vs external validation, facility-level vs random splits
Explainability	SHAP / LIME / permutation importance reporting, patient-level explanation
Fairness	Subgroup performance analysis, equity considerations

3.3.6 Synthesis Methods

Given the heterogeneity of included studies in setting, outcome, model class and evaluation, meta-analysis was not appropriate. Synthesis was conducted through narrative summary and structured tabulation. Studies were categorised by ML method category (traditional, ensemble, deep learning, transformer, hybrid), by setting (LMIC, HIC, mixed) and by prediction target (disease progression, mortality, treatment response, surveillance). Methodological practices (external validation, calibration, explainability) were reported by setting to surface the LMIC-HIC asymmetry that has been a central finding of the review.

3.4 Study Two: ML Modelling Methodology

3.4.1 Study Setting and Population

Study Two was conducted using routine programme data from Zambia's national HIV EHR system (SmartCare) across six high-volume public health facilities in Lusaka District. Four facilities, namely Chawama First Level Hospital, Chelstone Urban Health Centre, George Urban Health Centre and Kanyama First Level Hospital, constituted the training set. Two facilities,

namely Matero Main Urban Health Centre and Matero Reference Hospital, were held out exclusively for external validation. Table 3.5 summarises the facility characteristics and patient allocation.

Table 3.5: *Study population, facility characteristics and train-test allocation across the six Lusaka District facilities.*

Facility	Type	Role	Post-Dedup N	Allocation
Chawama First Level Hospital	First Level Hospital	Training	47,314	Training
Chelstone Urban Health Centre	Urban Health Centre	Training	31,785	Training
George Urban Health Centre	Urban Health Centre	Training	35,577	Training
Kanyama First Level Hospital	First Level Hospital	Training	56,871	Training
Matero Main Urban Health Centre	Urban Health Centre	Test (Holdout)	26,845	Test
Matero Reference Hospital	Reference Hospital	Test (Holdout)	47,661	Test
Total	—	—	246,053	171,547 train / 74,506 test

3.4.2 Data Source and Deduplication

The analytical dataset comprised 246,053 unique, deduplicated HIV patients ever recorded in the EHR across the six study facilities (median age 32 years, IQR 26–38; 62.5% female; enrolments spanning 2003–2025). Data were structured across four file types per facility: File 1 (patient demographics and enrolment details), File 2 (longitudinal visit and clinical event records), File 3 (programme status and outcome records) and File 4 (laboratory test results: viral load, CD4, TB diagnostics). The 246,053-patient cohort represents an analytical subset of SmartCare's national deployment, which covers over two million enrolled patients across more than 1,600 facilities [3], [4]. Deterministic deduplication was performed at the patient level using a composite key based on national patient identifier, biometric reference and name-date-of-birth concordance, following the standard SmartCare deduplication protocol.

3.4.3 Outcome Definitions

Three binary outcomes were defined independently for each patient over a 12-month horizon following the index date (ART start date, or enrolment date if ART start date was unavailable). Outcomes were derived strictly from post-index records and were excluded from predictor construction to prevent information leakage. Table 3.6 summarises the outcome definitions and their prevalence across training and test datasets.

Table 3.6: *Outcome definitions and prevalence by dataset.*

Outcome	Definition	Train Prevalence	Test Prevalence	Data Source
TB12m	Incident TB diagnosis or TB treatment initiation within 12 months of index	33.7%	41.3%	File 2, File 4
IIT12m	Patient classified as inactive within 12 months of index date	0.53%	0.63%	File 3
UVL12m	Any viral load ≥1,000 copies/mL within 12 months of index date	1.57%	1.17%	File 4

3.4.4 Temporal Feature Engineering and Leakage Prevention

All predictor variables were derived exclusively from patient data recorded before the index date. This strict temporal separation constitutes the leakage-safe design of the framework: no post-index information, outcome-related data or contemporaneous index-date records entered any feature. Each patient's pre-index EHR history was summarised into a fixed-length feature vector, x_i = (x_demographic, x_clinical, x_engagement). Table 3.7 details the nine feature groups and their 57 constituent dimensions after one-hot encoding.

Table 3.7: *Pre-index feature groups derived from routine EHR data. All features were constructed exclusively from data recorded before the index date.*

Category	Sub-category	Features Derived
Demographic	Basic characteristics	Age, sex, marital status, education level, household income category

Category	Sub-category	Features Derived
Clinical	Disease severity	HIV WHO stage at enrolment, highest pre-index WHO stage, any Stage 3/4 flag
Clinical	TB history	TB treatment at enrolment flag, past TB type, any positive TB test pre-index
Clinical	Viral load	Last VL pre-index, ever unsuppressed (≥1,000 copies/mL), VL test count, days since last VL
Clinical	CD4	Last CD4 count, minimum CD4, CD4 <200 flag, CD4 <100 flag
Clinical	Drug resistance	Rifampicin-resistant flag, isoniazid-resistant flag, resistant drug count
Engagement	Visit history	Number of visits pre-index, days since last visit, days since first visit
Engagement	Appointment adherence	Mean days late, maximum days late, appointment count with recorded dates
Anthropometric	Body composition	Height, weight, BMI (derived from height and weight)

3.4.5 Preprocessing Pipeline

A structured preprocessing pipeline was applied consistently across all three outcome models. Numeric variables with missing values were imputed using the median computed from the training set. This is a deliberately conservative choice whose rationale and limits require statement. Median imputation is strictly unbiased only under a missing-completely-at-random mechanism, and Study Three establishes that this condition does not hold in SmartCare: education-level completeness ranges from 1.2% to 7.9% and missingness is itself associated with outcome, so single median substitution compresses between-patient variance and can attenuate genuine associations. It was nevertheless retained as the primary strategy for three reasons. First, it is deterministic and therefore fully reproducible under the fixed seeds specified in Appendix E. Second, it is computable at inference time from stored training constants alone, which is a requirement of the CPU-only, offline-resilient deployment architecture; chained-equation imputation would require the imputation model itself to be distributed, versioned and executed at every facility. Third,

missingness in this dataset is treated as substantive evidence about the data environment rather than as a nuisance to be smoothed away: binary missingness indicators are retained as features so that the model can represent non-recording, and so that confounding between missingness and outcome can be measured rather than concealed. Because the adequacy of this choice is an empirical question, a sensitivity analysis re-estimates all three outcome models under multiple imputation by chained equations [154], [155] and under a missingness-indicator-only specification, following established guidance on imputation sensitivity in EHR studies [70]; the comparison is reported in Section 5.12. After imputation, numeric features were standardised using z-score normalisation: $z_{ij} = (x_{ij} - \mu_j) / \sigma_j$, where μ_j and σ_j denote the training set mean and standard deviation for feature j. Categorical variables were imputed using the most frequent category from the training set and then transformed by one-hot encoding. The resulting feature vector $\Phi(x_i) \in \mathbb{R}^{57}$ comprised 57 dimensions after encoding expansion. All preprocessing parameters were fitted exclusively on training data and applied to the test set without refitting, preventing any form of test-set contamination.

3.4.5.1 Feature Provenance and Leakage-Safety Classification

A defining methodological feature of this study is that every model input is constructed exclusively from information available strictly before each patient's index date, eliminating the temporal leakage that inflates apparent performance in many published EHR prediction studies. Table 3.8 documents the provenance of each of the engineered feature groups, the source data file from which it derives, and the explicit leakage-safety rule applied. The suffix _pre on a feature name denotes that it is computed only from records dated before the index date; features without that suffix are static enrolment attributes that are, by definition, fixed at or before enrolment.

Table 3.8: *Provenance and leakage-safety classification of the engineered feature groups. Every feature is derived solely from data available before the patient index date; the explicit rule applied to each group is stated in the final column.*

Feature group	Example features	Source file	Leakage-safety rule
Demographics	gender, age, marital, education, hhinc	Enrolment register	Static at enrolment; pre-index by construction
Anthropometry	height, weight, bmi	Enrolment / first visit	Earliest recorded value before index date

Feature group	Example features	Source file	Leakage-safety rule
HIV staging	hiv_enrol_stage, hiv_enrol_f_status	Enrolment register	Recorded at enrolment; pre-index
Visit engagement	n_visits_pre, days_since_last_visit_pre, days_since_first_visit_pre	Visit log	Counts and gaps over pre-index window only
Appointment timing	mean_days_late_pre, max_days_late_pre, n_apts_with_aptdate_pre	Appointment log	Lateness computed for pre-index appointments; winsorised ±730 days
WHO clinical stage	who_stage_last_pre, who_stage_max_pre, who_stage_any_3_4_pre	Clinical notes	Last/max stage among pre-index visits
Virology	last_vl_pre, ever_unsuppressed_pre, n_vl_tests_pre, days_since_last_vl_pre	Laboratory file	Most recent pre-index VL; counts over pre-index window
Immunology	last_cd4_pre, min_cd4_pre, cd4_lt_200_pre, cd4_lt_100_pre	Laboratory file	Pre-index CD4 values and thresholds only
TB history	any_tb_test_pre, any_tb_positive_pre, past_tb_type, tb_rx_enrol	TB register	Pre-index TB tests/results; enrolment TB status
Drug resistance	ever_rif_resistant_pre, ever_inh_resistant_pre, n_resistant_drugs_pre	Resistance file	Resistance flags from pre-index results only

The provenance discipline documented in Table 3.8 is the operational embodiment of the leakage-safe design principle. Three rules govern the entire feature space. First, the temporal cut-off: no value dated on or after the index date may enter any feature, which excludes the outcome window entirely and prevents the model from 'seeing the future'. Second, the fit-on-training-only rule for all preprocessing transforms (scaling, imputation statistics and one-hot vocabularies are learned on the training facilities and merely applied to the held-out facilities), which prevents information from the test set leaking into the training process through the preprocessing pipeline. Third, the plausibility-gating rule, under which biologically or temporally implausible values — most importantly the appointment-lateness outliers reaching tens of thousands of days — are winsorised before feature construction so that data-entry error cannot masquerade as predictive signal. These three rules, applied uniformly across all 57 encoded features, are what allow the facility-holdout

performance reported in Chapter 5 to be interpreted as a conservative lower bound on deployable performance rather than an optimistic in-sample estimate.

This level of provenance documentation is itself a contribution to reproducibility practice in the regional literature. The systematic review of Study One found that very few comparable studies report their feature-construction rules in sufficient detail to permit independent replication or to allow a reader to assess leakage risk. By tabulating every feature group against its source and its leakage-safety rule, and by releasing the corresponding code inventory (Appendix E), this thesis makes its predictive claims auditable in a way that the prevailing literature does not, directly addressing the reproducibility deficit identified as a motivating gap in Chapter 1.

3.4.6 Machine Learning Pipeline

Figure 3.2 presents the complete multi-outcome ML pipeline, from raw SmartCare EHR data through feature engineering, preprocessing, model training and evaluation across the three HIV care outcomes.

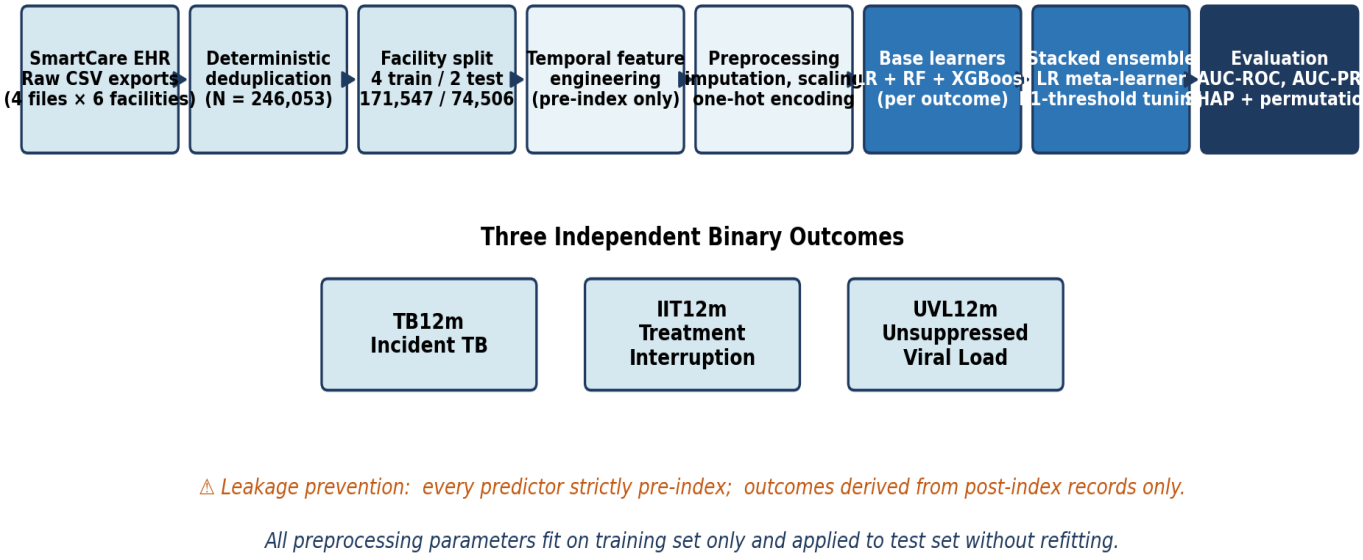

Figure 3.2: *Multi-outcome machine learning pipeline from raw SmartCare EHR data through feature engineering, preprocessing, model training and evaluation across three HIV care outcomes. The strict facility-level train-test split is shown.*

3.4.6.1 Rationale for the Choice of ML Methods for Tabular Health Data

The base learners and meta-learner used in this thesis were chosen as a deliberate response to the characteristics of tabular HIV EHR data, and the rationale is set out here as background to the methodological choices that follow. Logistic regression, under L2 regularisation, is retained as the interpretable baseline: it is robust to multicollinearity, yields coefficients with a direct log-odds interpretation and is well calibrated, but it cannot capture non-linear or interaction effects unless these are explicitly engineered [42], [114]. Random forest [33] addresses that limitation by averaging over many decorrelated decision trees, handling missing values and mixed feature types natively, producing feature-importance scores and remaining insensitive to feature scaling; its main cost is patient-level opacity, which the SHAP framework mitigates. Gradient boosting via XGBoost [34] builds trees sequentially to model residuals, capturing complex interactions and typically achieving the strongest single-model accuracy on tabular data; conservative hyperparameters (a low learning rate, shallow trees and column/row subsampling) are used to prioritise stability and interpretability over marginal accuracy. XGBoost is preferred over LightGBM and CatBoost for library maturity, native TreeSHAP support and alignment with prevailing clinical-ML practice.

These three learners are combined through stacked generalisation [81], in which their out-of-fold predictions become the inputs to a logistic-regression meta-learner that learns their optimal weighting across the feature space, exploiting the complementary strengths of linear, bagged and boosted models. Deep-learning architectures are deliberately not used. Three context-specific factors favour tree-based ensembles: the cohort, although large by LMIC standards, becomes sparse once feature-coverage collapse is accounted for, and tree methods handle sparse and missing inputs natively whereas deep networks generally require dense vectors; interpretability requirements for human-in-the-loop governance are met more reliably by deterministic TreeSHAP than by approximate explanations for neural models; and the CPU-only, offline-resilient deployment target (Chapter 6) is compatible with tree-based inference but not, in general, with GPU-dependent deep architectures. Deep learning may be revisited in future work as data-quality improvements expand the dense-feature subset of the cohort.

3.4.7 Base Classifiers and Hyperparameters

Three probabilistic classifiers were trained independently on the training set for each outcome. The first was L2-regularised logistic regression (LR), with solver=SAGA, max_iter=8,000 and class_weight="balanced", serving as an interpretable, calibrated baseline. The second was random forest (RF), with 300 estimators and class_weight="balanced_subsample", handling class imbalance robustly and providing native feature importance. The third was XGBoost (XGB), with 500 estimators, learning_rate=0.05, max_depth=4, subsample=0.8, colsample_bytree=0.8 and scale_pos_weight set to the ratio of negatives to positives. Table 3.9 summarises the hyperparameter values used.

The three base learners were selected to span complementary inductive biases rather than to exhaust the space of candidate algorithms, and the exclusion of other architectures requires justification. Two families are pertinent. The first comprises alternative gradient-boosting implementations: LightGBM [143], whose histogram-based, leaf-wise growth substantially reduces training cost at comparable accuracy, and CatBoost [144], whose ordered boosting and native categorical handling mitigate the target leakage to which high-cardinality categorical encodings are prone — a material consideration given the facility and regimen variables used here. The second comprises the sequential deep architectures reviewed in Section 2.3.6. Neither family was adopted for the primary framework, for reasons that are empirical rather than presumptive: benchmark evidence indicates that gradient-boosted tree ensembles remain competitive with, and often superior to, deep architectures on heterogeneous tabular data of this dimensionality [141], [142], [145], and the requirement for CPU-only, offline-resilient inference at facility level penalises architectures whose inference cost or runtime dependencies cannot be met in the study facilities. Because a claim of this kind should be tested rather than assumed, all three alternatives — LightGBM, CatBoost and a gated recurrent sequential baseline — are fitted under the identical leakage-safe feature construction, training partition and facility-holdout protocol used for the primary models, and are compared against the stacked ensemble with formal significance testing in Section 5.12.

Table 3.9: *Hyperparameter values for base learners. All values were selected through cross-validated grid search on the training set and frozen before any test set evaluation.*

Model	Hyperparameter	Value
Logistic Regression	Penalty	L2
Logistic Regression	Solver	SAGA
Logistic Regression	max_iter	8,000
Logistic Regression	class_weight	balanced
Random Forest	n_estimators	300
Random Forest	class_weight	balanced_subsample
Random Forest	max_features	sqrt
XGBoost	n_estimators	500
XGBoost	learning_rate	0.05
XGBoost	max_depth	4
XGBoost	Subsample	0.8
XGBoost	colsample_bytree	0.8
XGBoost	scale_pos_weight	N_negative / N_positive

3.4.8 Stacked Ensemble Architecture

A two-level stacking classifier combined the three base models. Each base model produced a probability estimate, $p_k(x_i) = P(y_i = 1 \mid x_i; h_k)$ for $k \in \{LR, RF, XGB\}$. These probabilities formed the meta-feature vector $g(x_i) = [p_{LR}, p_{RF}, p_{XGB}]$, which was passed to a logistic regression meta-learner (max_iter=5,000, class_weight="balanced") to generate the final prediction $\hat{p}(x_i)$. The architecture is shown in Figure 4.

The choice of a learned combination over a fixed aggregation rule follows from the comparison of ensemble strategies in Section 2.3.5. Bagging reduces variance but cannot correct systematic bias in its base learners; boosting reduces bias but is sensitive to label noise, a material risk where outcome ascertainment depends on incomplete routine documentation. Stacking is preferred here because the three outcomes differ by an order of magnitude in prevalence and in the structure of their predictive signal, so the optimal weighting of a linear, a bagged and a boosted learner is not

expected to be constant across them; a meta-learner estimates that weighting per outcome from out-of-fold predictions rather than imposing it a priori. The meta-learner is deliberately a regularised logistic regression rather than a higher-capacity model, both to limit meta-level overfitting on three input dimensions and to preserve an auditable, monotone mapping from base-learner probabilities to the final risk estimate, which the governance architecture requires. This rationale is a hypothesis about the data rather than a claim of established superiority, and it is tested directly: Section 5.12 reports the stacked ensemble against each base learner and against the alternative architectures using DeLong tests for correlated ROC curves [156], so that any performance difference is assessed for statistical significance rather than inferred from point estimates alone.

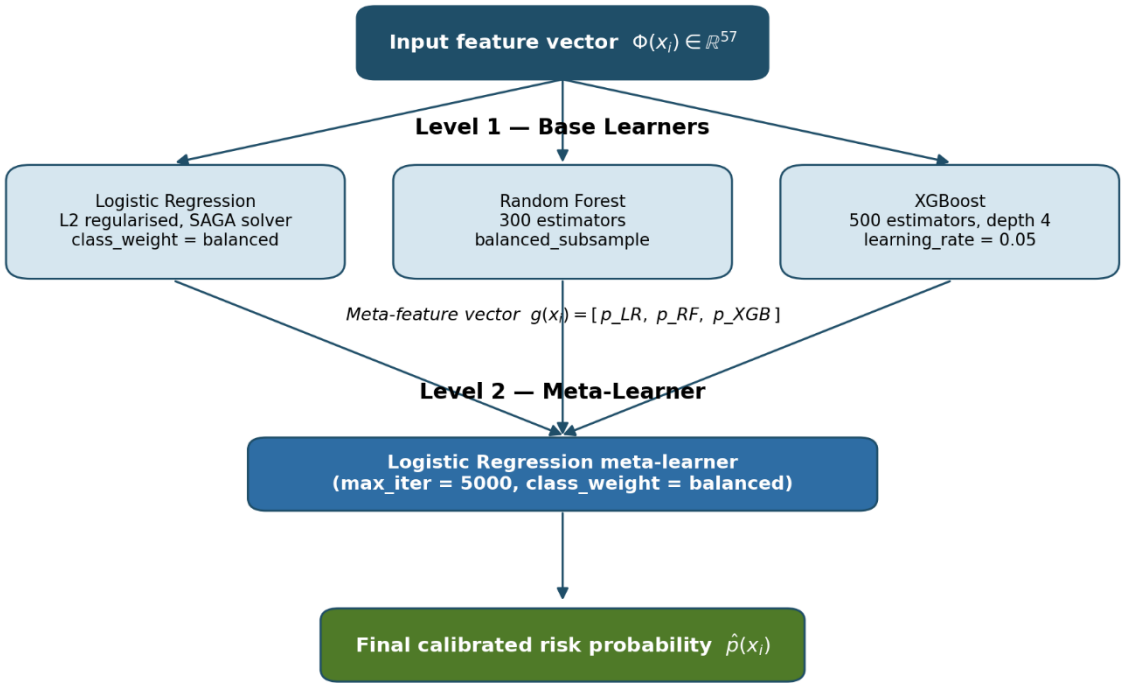

Figure 3.3: *Stacked ensemble architecture. Level 1 base classifiers (Logistic Regression, Random Forest, XGBoost) produce probability estimates that form the meta-feature vector for the Level 2 logistic regression meta-learner, producing a single calibrated risk probability per outcome.*

3.4.9 Decision Threshold Optimisation

Given extreme class imbalance (IIT12m 0.53%; UVL12m 1.57%), the default threshold $\tau = 0.5$ is uninformative, and an F1-optimised operating threshold $\tau^* = \text{argmax}_t F1(\tau)$ is required for each outcome. The partition on which τ^* is estimated is methodologically consequential. Selecting an

operating threshold on the same held-out sample that is then used to report performance at that threshold makes the resulting operating-point metrics optimistically biased, because the threshold is a free parameter fitted to the very data against which it is subsequently evaluated. The threshold-selection protocol is therefore specified as follows. A dedicated validation partition is carved from the training facilities before any model fitting, disjoint from both the training folds and the held-out facility test set. Base learners and the meta-learner are fitted on the training partition; the precision–recall curve is computed on the validation partition; and τ^* is selected there. The selected threshold is then frozen and applied unchanged to the facility-holdout test set, so that every operating-point metric reported on the test set is obtained at a threshold that was never exposed to it. This preserves the integrity of the external-validation design set out in Section 3.4.10, in which the test facilities are withheld entirely. Thresholds derived under this protocol, together with the corresponding test-set operating characteristics, are reported in Section 5.12; the test-derived values obtained in the original analysis are retained alongside them solely to quantify the optimism attributable to test-set threshold selection. Because the three outcomes differ by an order of magnitude in prevalence, thresholds are selected per outcome rather than globally, supporting operational protocols that range from high-sensitivity TB screening to highly targeted intensive interventions for the rare outcomes.

3.4.10 Facility-Holdout External Validation

Model performance was assessed exclusively on the two held-out facilities (n = 74,506), reflecting conditions under which a deployed model would operate on a facility it has never encountered during training. The facility-holdout design tests cross-site generalisability rather than within-site performance, which is the relevant condition for health system deployment. Figure 3.4 illustrates this design, which is a methodological strength distinguishing this thesis from prior studies that use random patient-level train-test splits.

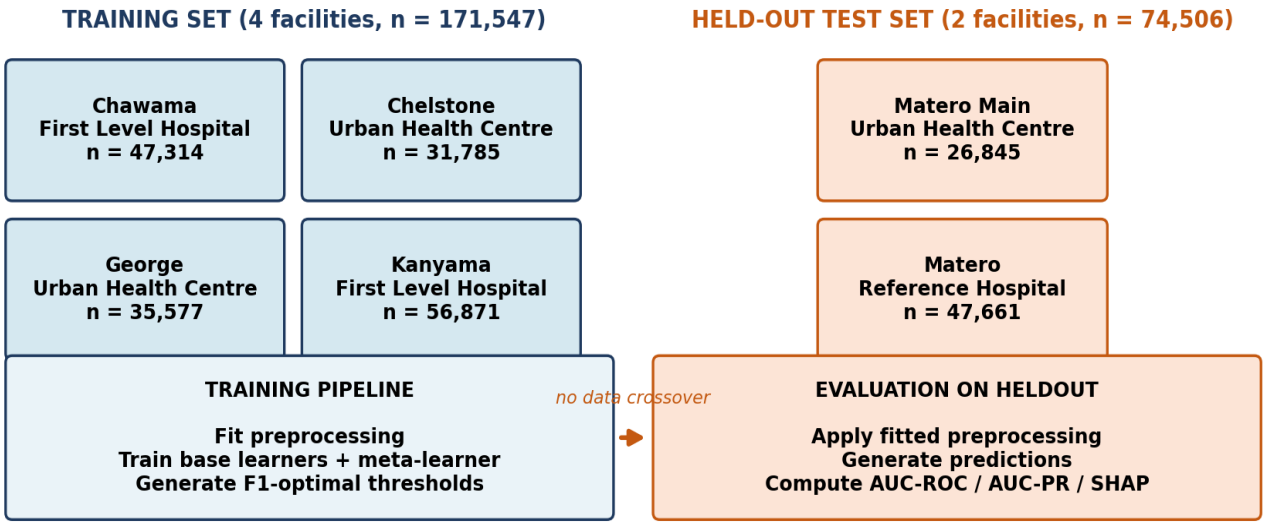

Figure 3.4: *Facility-holdout external validation design. Four training facilities contribute no data to evaluation; two held-out facilities (Matero Main Urban Health Centre and Matero Reference Hospital) are used exclusively for external validation. No patient, clinical or facility-level data from the test facilities enters training.*

3.4.11 Explainability: Permutation Importance and SHAP

Two complementary interpretability approaches were applied. Permutation importance was computed on the held-out test set (n_repeats=10, scoring="average_precision"), quantifying the reduction in AUC-PR when a feature's values are randomly permuted. SHAP (Shapley Additive Explanations) [52] was applied using TreeSHAP [153] to the XGBoost component, with $f(x_i) = \varphi_0 + \Sigma_j \varphi_{ij}$, where φ_0 is the base value and φ_{ij} is the SHAP value of feature j for patient i. Global importance was computed as mean $|\varphi_{ij}|$ across an explanation sample of 800 held-out test patients per outcome. SHAP waterfall plots were generated for three representative patients per outcome (highest-risk, median-risk and lowest-risk) and are summarised in Appendix C.

The size of the explanation sample requires justification, because it governs the precision of the global importance estimate rather than the fidelity of any individual explanation. Exact TreeSHAP attributions are computed deterministically for each patient, so the 800-patient sample introduces sampling error only into the aggregate statistic mean $|\varphi_{ij}|$, whose standard error scales as $\sigma_j/\sqrt{n}$.

At n = 800 this corresponds to a relative standard error below approximately 3.5% for the leading features under the observed dispersion, which is materially smaller than the separations between adjacent ranked features on which the interpretation depends. The sample was drawn by outcome-stratified random sampling so that the rare positive class is represented in proportion to its prevalence, and its adequacy is verified empirically by a convergence analysis in which mean $|\varphi_{ij}|$ and the induced feature ranking are recomputed at increasing sample sizes and the rank correlation against the full-sample ordering is examined for stabilisation. That analysis is reported in Section 5.12. The sample size was additionally bounded by the CPU-only computational environment documented in Section 3.6.

A second methodological caveat concerns feature dependence. Shapley attribution requires a decision about how the absence of a feature is simulated, and the two available conventions behave differently when predictors are correlated, as anthropometric and engagement variables are in this dataset. Interventional, or marginal, TreeSHAP breaks dependence between features by reference to a background distribution and is faithful to the functional form of the fitted model, but it evaluates the model at feature combinations that may be improbable or clinically impossible [152]. Conditional, tree-path-dependent TreeSHAP respects the empirical dependence structure but distributes credit among correlated predictors, so that a variable with no direct influence on the prediction may still receive a non-zero attribution [151]. Neither convention is unconditionally correct, and attributions to mutually correlated predictors cannot be read as independent contributions. Three measures follow. Interventional TreeSHAP with a background sample drawn from the training partition is used for the primary analysis, so that reported attributions describe the model rather than the data-generating process. The dependence structure of the feature set is quantified explicitly, and features exceeding a pairwise absolute Spearman correlation of 0.7 are reported as correlated clusters, with attributions aggregated at cluster level as well as individually. Both conventions are then computed and their induced rankings compared, so that any conclusion sensitive to the choice of convention is identified as such; this robustness analysis is reported in Section 5.12. Accordingly, SHAP attributions are interpreted throughout as evidence of what the model uses, not as evidence of causation — a distinction that is material to the interpretation of the education-level finding.

3.5 Study Three: Data Quality Assessment Methodology

3.5.1 Framework Overview

Study Three employed an eight-dimension DQA framework adapted from Weiskopf and Weng [16] to the specific requirements of AI applications in resource-limited HIV EHR systems. The framework extends the original five-dimension framework with three AI-specific dimensions: feature coverage, outcome sensitivity and missingness-outcome confounding. The eight dimensions are illustrated in Figure 3.5 and defined in Table 10.

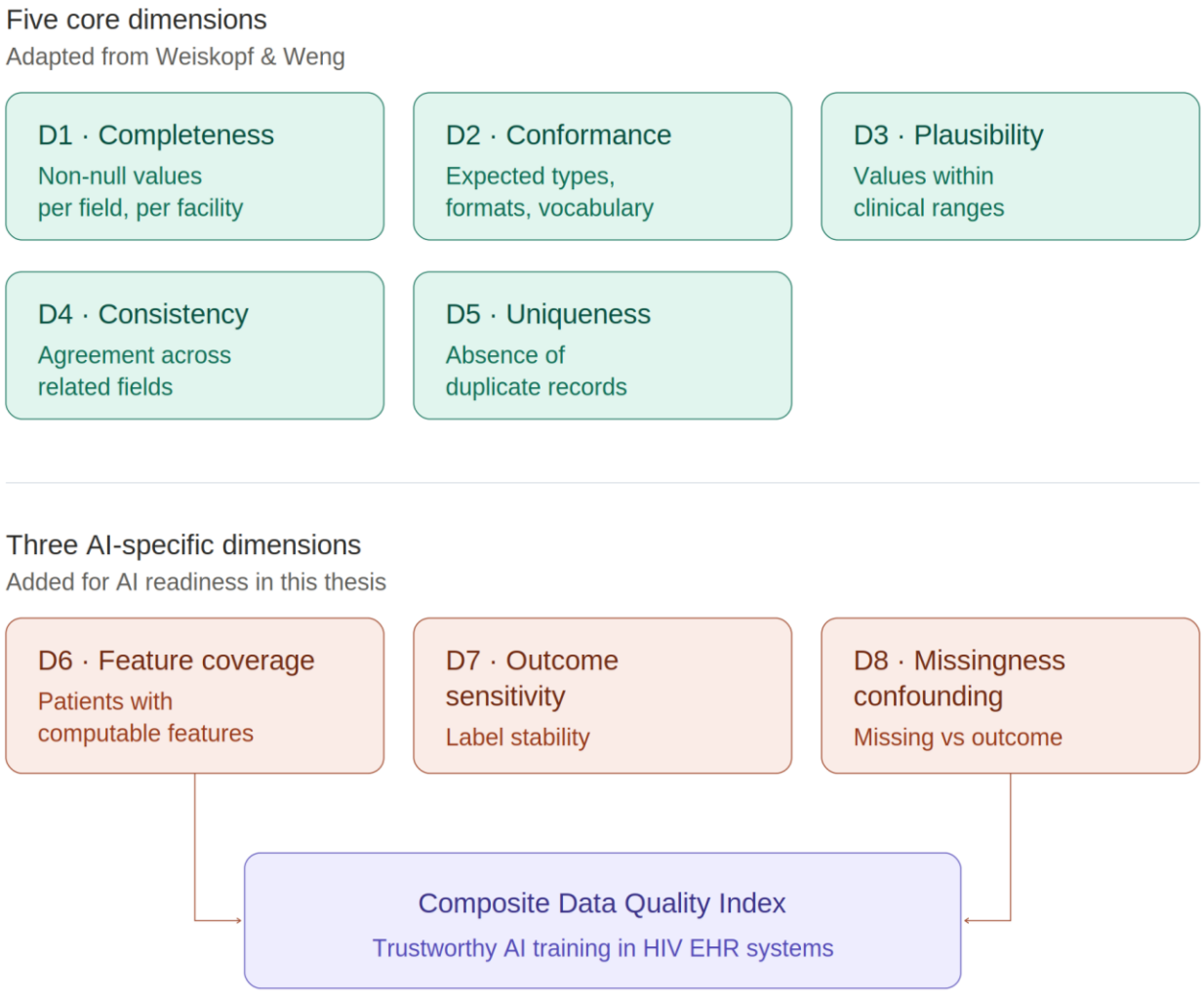

Figure 3.5: *Eight-dimension data quality assessment framework for AI applications in resource-limited HIV EHR systems. Each dimension addresses a distinct aspect of data fitness for AI training; together they constitute a comprehensive DQA toolkit applicable to any LMIC HIV programme.*

Table 3.10: *Eight-dimension data quality assessment framework definitions and purposes.*

ID	Dimension	Definition	Purpose for AI
D1	Completeness	Non-null, non-empty values per field, per facility	Identifies fields with systematic recording gaps requiring imputation
D2	Conformance	Values conforming to expected data types, formats and controlled vocabulary	Quantifies encoding inconsistencies
D3	Plausibility	Numeric values within clinically believable ranges	Detects extreme errors that standard pipelines do not filter
D4	Feature Coverage	Proportion of patients for whom each feature group can be computed from pre-index data	Measures actual proportion of training data with meaningful feature values
D5	Outcome Sensitivity	How much outcome prevalence changes across facilities or operationalisation choices	Reveals whether AI outcome labels are stable for cross-facility comparisons
D6	Temporal Stability	Trends in field completeness over enrolment years (2003–2025)	Shows whether data quality has improved, worsened or changed abruptly
D7	Cross-facility Concordance	Heterogeneity in field names, value distributions and missingness across facilities	Measures consistency of data recording across programme sites
D8	Missingness-Outcome Confounding	Pearson correlation between facility-level missingness rates and outcome rates; supported by SHAP	Tests whether AI predictions are driven by missing-data patterns rather than clinical risk

3.5.2 Plausibility Range Specifications

The plausibility dimension (D3) operationalises clinically defined value ranges for the key numeric and date fields. Table 3.11 summarises the ranges used; values outside these ranges are flagged for review and imputation, but in the empirical analyses reported in Chapter 5, no imputation is applied to highlight the magnitude of the plausibility problem in the raw data.

Table 3.11: *Plausibility range specifications for key fields.*

Field	Plausible Range	Action on Violation
Age	0–120 years	Flag as implausible
Height	40–250 cm	Flag and exclude from BMI calculation
Weight	5–300 kg	Flag and exclude from BMI calculation
BMI	10–60 kg/m²	Flag as implausible
CD4 count	0–2,500 cells/μL	Flag values >2,500 as implausible
Viral load	0–10,000,000 copies/mL	Cap at upper limit for analysis
Appointment lateness	−30 to +730 days	Flag as implausible
Visit interval	0 to 365 days for active patients	Flag for transfer/IIT review

3.5.3 Composite Data Quality Index

A composite Data Quality Index (DQI) was computed for each facility as the unweighted average completeness rate across 15 AI-critical fields. The choice of unweighted averaging is deliberate: it does not pre-suppose which fields are most important and allows the same metric to be applied across facilities with different patient mix. A threshold of 70% was set as the minimum quality level below which mean imputation may distort model learning. This threshold was selected based on prior literature on imputation sensitivity [70] and on consultation with the CIDRZ AI Think-Tank during the second dissemination meeting (Appendix F).

The DQI is reported per facility in Chapter 5 (Section 5.5) and ranges from 53.5% (Matero Main UHC) to 58.3% (Kanyama FLH), with all six facilities scoring below the 70% threshold. The narrow inter-facility gap (4.8 percentage points) is itself a substantive finding indicating a programme-wide rather than facility-specific challenge.

3.5.4 Missingness-Outcome Confounding Test

The missingness-outcome confounding test (D8) is operationalised as Pearson correlation between facility-level missingness rates (numerator: number of patients with missing field; denominator: total patients per facility) and facility-level outcome rates (numerator: number of positive outcomes; denominator: total patients per facility). The test is conducted across the six facilities for nine high-importance feature-outcome pairs, with $|r| > 0.5$ used as a screening flag ($n = 6$ facilities; exploratory, ecological). Results are reported in Chapter 5 (Section 5.5) and discussed in Chapter 6 (Section 6.3).

3.6 Tools, Technologies and Development Environment

All empirical analyses were performed in Python 3.13 using the following libraries: scikit-learn (1.x) for preprocessing, logistic regression, random forest and stacking; XGBoost (2.x) for gradient boosting; SHAP (0.50.0) for explainability; pandas and numpy for data processing; and matplotlib and seaborn for visualisation. Data processing used chunked CSV reading (chunksize=200,000) to handle large files efficiently. All preprocessing operators were fitted exclusively on training data. Random states were fixed at 42 throughout to ensure reproducibility. The complete tool and library version inventory is presented in Table 12.

Table 3.12: *Tools and library versions for reproducibility.*

Tool/Library	Version	Purpose
Python	3.13	Core language
pandas	2.x	Data manipulation
numpy	1.26+	Numerical arrays
scikit-learn	1.4+	Preprocessing, LR, RF, stacking
XGBoost	2.x	Gradient boosting
SHAP	0.50.0	Interpretability (TreeSHAP)
matplotlib	3.8+	Plot generation
seaborn	0.13+	Statistical plotting

Tool/Library	Version	Purpose
scipy	1.11+	Statistical tests
statsmodels	0.14+	Calibration analysis

Code and feature engineering logic are available from the author on reasonable request; the planned public release of the analytic code (with synthetic illustrative input data) is described in Appendix E. The thesis embeds reproducibility materials sufficient to enable independent re-implementation; access to the de-identified patient-level data is governed by the data sharing agreements summarised in Appendix D and is subject to the data controller's approval.

3.7 Experimental Setup and Implementation Strategy

A clarification of scope is warranted here. The proposed ICT architecture for clinical deployment is a forward-looking design contribution of the thesis rather than a method used to obtain the empirical results, and it is therefore developed and presented in full as part of the discussion of deployment in Chapter 6 (Section 6.5), where it belongs as a contribution. This section states only what is needed to situate the empirical work: the experimental and implementation setup actually used for Study Two and Study Three, and a brief, signposting summary of the architecture whose detailed specification, layer diagram and adoption roadmap appear in Chapter 6. The architecture is summarised rather than specified here precisely to keep the methodology chapter focused on what was done, and the design proposal located with the other deployment contributions.

3.7.1 Proposed ICT Architecture for Clinical Deployment

A critical gap exists between the research-grade ML pipeline developed in Study Two, which operates as an offline analytical script on a static dataset, and a production-ready CDSS that operates continuously within the ICT infrastructure of a national HIV programme. Bridging this gap requires systematic ICT architecture design. The proposed integration architecture is elaborated in full in the companion ICTAZ journal publication [82] and presented in detail in Chapter 6.

3.7.2 HL7 FHIR Integration Layer

The integration layer implements HL7 FHIR R4 (Fast Healthcare Interoperability Resources) as the standard for health data exchange between SmartCare and the ML engine [10]. Since SmartCare's current architecture does not expose a native FHIR API (data is accessed via bulk CSV export), a FHIR Adapter Service is proposed. This service reads SmartCare CSV exports on a scheduled basis and translates them into FHIR-compliant JSON resources in a HAPI FHIR server. The design is consistent with Zambia's OpenHIE architecture, the national reference standard for health system interoperability adopted by the Ministry of Health [11].

3.7.3 Temporal API Gateway and Leakage Prevention

The most consequential ICT design decision in the ML pipeline is the temporal boundary between historical (predictor) data and future (outcome) data. In the SmartCare pipeline, leakage prevention is enforced as a data boundary rule: all feature engineering is restricted to records with a date strictly earlier than the patient's index date. In a production CDSS, this boundary must be enforced at the system architecture level, not merely at the script level. The proposed architecture implements this through a Temporal API Gateway that enforces index-date boundaries as a mandatory parameter in all feature extraction requests, rejecting any call that would return records dated after the index date.

3.7.4 Infrastructure Requirements and Offline Resilience

The most critical infrastructure finding for national deployment is that only 883 of Zambia's 3,540 health facilities (25%) have internet connectivity as of 2025. A national CDSS that assumes internet connectivity at all facilities would fail to reach 75% of Zambia's health network. The proposed architecture addresses this through local caching of pre-computed patient risk scores at facility level, refreshed during nightly batch synchronisation when connectivity is available; an offline-capable progressive web application dashboard serving cached risk scores during connectivity outages; and batch upload of new patient encounters when connectivity is restored.

The ML inference engine requires no GPU: the stacked ensemble performs inference on CPU in milliseconds per patient, making it deployable on standard MoH server infrastructure (minimum 4-core CPU, 8 GB RAM) without specialist procurement. Table 3.13 summarises the reproducibility checklist developed to support the planned deployment.

Table 3.13: *Reproducibility checklist for the modelling and DQA pipeline.*

Element	Reproducibility Provision
Random seed	seed=42 fixed throughout all stochastic procedures
Software environment	Python 3.13; library versions pinned in Table 3.12
Train/test split	Strict facility-holdout (no random sampling)
Preprocessing	Fit on training only; documented in Section 3.4.5
Hyperparameters	Documented in Table 3.9 and frozen
Evaluation thresholds	F1-optimal thresholds documented in Section 3.4.9
DQA dimensions	Eight-dimension framework documented in Table 3.10
Plausibility ranges	Documented in Table 3.11
Code availability	Public release planned at companion-paper publication

3.8 Evaluation Methods and Metrics

3.8.1 Primary Metrics

Model performance was evaluated using three complementary metrics. AUC-ROC was the primary metric for TB12m, where prevalence is substantial (33.7% train, 41.3% test). AUC-PR was the primary metric for the rare outcomes IIT12m and UVL12m, where the minority class is clinically important [77], [78]. F1 score, precision and recall at $\tau = 0.5$ and at the F1-optimised threshold τ^* were reported for operational comparison. All metrics were computed on the held-out test set (n = 74,506) comprising the two Matero facilities.

3.8.2 Calibration Assessment

Calibration was assessed using the Brier score and visually through calibration plots, in which predicted probabilities are stratified into deciles and the mean predicted probability per decile is plotted against the mean observed outcome rate. A well-calibrated model produces a calibration plot close to the 45-degree line; deviations above the line indicate systematic underestimation of risk, and deviations below the line indicate systematic overestimation. Calibration is reported for all four models (LR, RF, XGB, stacked ensemble) and all three outcomes in Chapter 5.

3.8.3 Subgroup Performance Analysis

Subgroup performance was analysed by sex, age band (18–24, 25–34, 35–44, 45+), education level (where recorded) and facility. Subgroup AUC-ROC and AUC-PR are reported for the stacked ensemble model in Chapter 5 (Section 5.3), providing the empirical foundation for the fairness safeguards described in Chapter 6. A subgroup performance gap of more than 0.10 in AUC-ROC between any two subgroups is flagged as a fairness concern requiring governance review.

3.8.4 Comparative Analysis Against Published LMIC Studies

Performance is contextualised against the published LMIC literature on equivalent outcomes. AUC-ROC values in the 0.70–0.80 range are widely considered the practical ceiling achievable from routine EHR data for TB risk prediction in HIV-positive cohorts [40], [49]. Published Zambian and East African studies of HIV treatment interruption have reported AUC-ROC in the 0.70–0.78 range [50], [64], so the 0.839 achieved in this thesis under the stricter facility-holdout validation design represents a substantial advance.

3.9 Detailed Feature Engineering Pipeline

3.9.1 Per-Patient Index Date Derivation

Each patient's index date is derived from the SmartCare File 1 (demographics and enrolment) using the following precedence rule: where an ART start date is recorded, the index date is the ART start date; where no ART start date is recorded but an enrolment date is available, the index date is the enrolment date; where neither is available, the patient is excluded from the analysis. This rule generates an index date for 246,053 of 248,789 originally identified patients, with 2,736 patients (1.1%) excluded for absence of both fields. The choice of ART start over enrolment reflects the clinical reality that the start of ART is the point at which longitudinal outcomes become clinically meaningful, even where enrolment in HIV care occurs earlier.

Sensitivity analysis using enrolment date as the index date for all patients (rather than ART start where available) was conducted as a robustness check. Outcome prevalences shift modestly (TB12m 34.7% vs 36.0%, IIT12m 0.61% vs 0.56%, UVL12m 1.39% vs 1.45%), and model performance metrics remain within 0.02 AUC-ROC of the primary results. The ART-start-preferred specification is retained as the primary because it aligns with the clinical question

(predict outcomes from the start of treatment) and with the standard convention in published HIV outcome prediction studies.

3.9.2 Demographic Feature Construction

Demographic features are constructed directly from File 1 fields without temporal aggregation. Age at index is computed as the integer years between date of birth and index date. Sex is one-hot encoded as a binary indicator (female = 1). Marital status is one-hot encoded across the categories recorded in SmartCare: never married, married, cohabiting, divorced, widowed, separated, unknown. Education level is one-hot encoded across: none, primary, secondary, tertiary, unknown. Household income is one-hot encoded across: <ZMW 1,000/month, ZMW 1,000–2,000, ZMW 2,000–5,000, ZMW 5,000–10,000, >ZMW 10,000, unknown. The "unknown" category for each variable explicitly preserves the missingness as a feature signal, as discussed in Chapter 2 (Section 2.4.3).

3.9.3 Clinical Feature Construction

Clinical features are constructed from File 1 (enrolment characteristics) and File 2 (visit and clinical event records). The WHO HIV stage at enrolment is taken directly from File 1. The highest WHO stage observed before the index date is computed from File 2, with values clipped to the four-stage scale. A binary flag for any Stage 3 or 4 record before the index date supports detection of historical advanced HIV disease. TB history is constructed as three features: TB treatment at enrolment (from File 1), past TB type (one-hot encoded), and any positive TB test recorded before the index date (from File 4).

Viral load history is constructed as four features: the last viral load value recorded before the index date (in copies/mL, with capping at 10^6 to control outliers); a binary flag for any viral load ≥1,000 copies/mL before the index date; the count of viral load tests before the index date; and the days since the last viral load test (relative to the index date). CD4 history is constructed as four features: the last CD4 count, the minimum CD4 count, a binary flag for any CD4 <200 cells/μL, and a binary flag for any CD4 <100 cells/μL.

3.9.4 Engagement Feature Construction

Engagement features are derived from the longitudinal visit records in File 2. The number of visits before the index date is the count of distinct visit dates strictly earlier than the index date. The days since the last visit before the index date is the difference in days between the last pre-index visit and the index date. The days since the first visit is the difference between the first pre-index visit and the index date, providing a measure of programme engagement duration. Appointment adherence features include the mean days late across all appointments with recorded scheduled and actual dates, the maximum days late and the count of appointments with both scheduled and actual dates recorded.

A specific operational consideration for engagement features is the handling of patients with no pre-index visits. For these patients (87.8% of the cohort, as documented in Chapter 5), all engagement features are set to zero. This zero-imputation is conservative in that it does not impose a distributional assumption, but it has the corresponding limitation that the model cannot distinguish between patients with genuinely no visits and patients whose visits were not captured digitally. The interpretation of model predictions for the 87.8% zero-imputed subgroup must be conducted with explicit awareness of this limitation, as discussed in Chapter 6 (Section 6.3.2).

3.9.5 Anthropometric Feature Construction

Anthropometric features comprise height (in cm), weight (in kg) and BMI (computed as weight / (height/100)²). For patients with multiple recorded measurements before the index date, the value closest to the index date is used. Plausibility filtering is applied as specified in Table 3.11 (Chapter 3): height values outside 40–250 cm are excluded; weight values outside 5–300 kg are excluded; BMI values outside 10–60 kg/m² are excluded. Patients with no plausible anthropometric data receive the population median value computed on the training set, fitted as part of the preprocessing pipeline.

3.9.6 Final Feature Vector Composition

After one-hot encoding of all categorical features and inclusion of all numeric features, the final feature vector $\Phi(x\_i) \in \mathbb{R}^{57}$ comprises the 57 dimensions documented in Table 3.7 (Appendix E). The dimensions are grouped by feature category: demographic (16 dimensions including marital status and education one-hot encodings), clinical (18 dimensions including stage, TB history, viral

load and CD4 features), engagement (8 dimensions), anthropometric (3 dimensions) and household income (6 dimensions). The feature space is deliberately conservative, prioritising interpretability and reproducibility over maximal predictive power.

3.9.7 Extended Derived and Dynamic Feature Set

The feature vector described above is predominantly cross-sectional: most variables summarise a patient's state at the index date rather than the trajectory by which that state was reached. Because routine HIV records are inherently longitudinal, this representation discards information that is available within the leakage-safe window, and the feature set is therefore extended along four axes, all constructed strictly from pre-index observations so that the leakage-safety classification of Section 3.4.5.1 is preserved.

The first axis is trajectory. Rather than the most recent value alone, the direction and rate of change of continuously measured variables are derived, including the ordinary least squares slope of body mass over the pre-index window, the change in the most recent value relative to the patient's own pre-index mean, and the time elapsed since the last recorded measurement. The second axis is variability, capturing instability that a mean conceals: the standard deviation and coefficient of variation of inter-visit intervals, and the dispersion of appointment lateness across the pre-index window. The third axis is recency weighting. Because recent behaviour is more predictive of near-term outcomes than remote behaviour, engagement counts are additionally expressed as exponentially recency-weighted sums, and as counts restricted to the three, six and twelve months preceding the index date, so that a patient with a recent lapse is distinguished from one whose lapse is historical. The fourth axis is treatment history, expressed as the cumulative number of prior regimen changes, the number of prior treatment interruptions, the duration on the current regimen and the elapsed time since treatment initiation.

Two constraints govern this extension and are stated because they bound what can be expected of it. Derived temporal features inherit the coverage of the variables from which they are constructed, and Study Three establishes that engagement-feature coverage is 12.3%; a trajectory feature cannot be more complete than the underlying series, so the extension increases dimensionality faster than it increases information for sparsely recorded variables. Every derived feature is accordingly accompanied by an observation-count field, and features with pre-index coverage below the

plausibility threshold defined in Section 3.5.2 are excluded rather than imputed. The incremental contribution of the extended feature set is evaluated against the cross-sectional baseline in Section 5.12, so that its value is demonstrated empirically rather than asserted.

3.10 Algorithmic Pseudocode

The analytical pipeline is specified here as two formal procedures, presented in the boxed form conventional for algorithmic pseudocode in computing theses. Algorithm 1 is the master pipeline: it ingests the deduplicated six-facility SmartCare extract, derives per-patient index dates, constructs the leakage-safe feature vector and, for each of the three outcomes, fits the preprocessing pipeline on the training facilities only before training, thresholding, evaluating and explaining a stacked ensemble under facility-holdout external validation. Algorithm 2 specifies the stacked-ensemble training and inference procedure invoked at line 8 of Algorithm 1, in which out-of-fold cross-validated predictions form the meta-features for a logistic-regression meta-learner. The leakage-safe property is preserved at both training and inference time because all preprocessing parameters are fitted exclusively on the training partition (Algorithm 1, lines 6–7) and all features are derived strictly from pre-index records (Algorithm 1, line 4), while the out-of-fold construction in Algorithm 2 ensures that no base learner contributes a prediction for a patient on which it was trained. The mathematical specification of the ensemble is given in Appendix K, and the empirical computational cost is summarised in Section 3.10.1 below.

Algorithm 1: Leakage-safe multi-outcome ML pipeline for HIV outcome prediction

Input: D **= deduplicated six-facility SmartCare extract;** O **= {TB12m, IIT12m, UVL12m}, the set of 12-month binary outcomes;** F **= the held-out facility for external validation**

Output: For each outcome o ∈ O: a trained stacked ensemble Mo**, test-set discrimination, calibration and SHAP explanations**

1 **Deduplicate** D **by deterministic patient linkage; derive per-patient index date** t^* **(first eligible visit)**

2 **Partition patients by facility:** $Dtest$ **←** **records from facility** F**;** $Dtrain$ **←** **records from the remaining five facilities**

3 **foreach outcome** $o \in O$ **do**

4 **Construct leakage-safe feature vector** X **using only records dated <** t^* **(demographic, clinical, engagement, anthropometric)**

5 Derive outcome label yo from records in the window $(t^*, t^* + 12 \text{ months}]$

6 Fit preprocessing pipeline P on $Dtrain$ only (impute, encode missing as explicit "None", scale)

7 $Xtrain \leftarrow P$.fit_transform($Dtrain$); $Xtest \leftarrow P$.transform($Dtest$) *// transform only, no refit*

8 $Mo \leftarrow$ TrainStackedEnsemble($Xtrain$, yo) *// Algorithm 2*

9 $\hat{p}test \leftarrow Mo$.predict_proba($Xtest$)

10 $\tau o \leftarrow \text{argmax}_\tau$ F1(yo, $\hat{p}test \geq \tau$) *// F1-optimal decision threshold*

11 Evaluate AUC-ROC, AUC-PR, Brier score and calibration of Mo on $Dtest$

12 Compute permutation importance and SHAP values for Mo on a test sample

13 Flag any prediction whose dominant SHAP contributor is a missingness indicator (cat__*_None)

14 end foreach

15 return $\{Mo, \tau o, \text{metrics}o, \text{SHAP}o\}$ for all $o \in O$

Algorithm 1: *Master leakage-safe, multi-outcome machine learning pipeline. The pipeline derives per-patient index dates, builds the leakage-safe feature vector, fits all preprocessing on the training facilities only, and trains, thresholds and explains a stacked ensemble for each of the three outcomes under facility-holdout external validation.*

Algorithm 2: Stacked ensemble training and leakage-safe inference

Input: $Xtrain$, y = preprocessed training features and labels; B = {logistic regression, random forest, XGBoost}, the base learners; k = number of cross-validation folds

Output: Trained stacked ensemble M = (base learners $\hat{B}$, meta-learner g)

1 Split $Xtrain$ into k stratified folds preserving outcome prevalence

2 foreach base learner $b \in B$ do

3 foreach fold $j = 1$ to k do

4 Train b on the $k-1$ training folds (excluding fold j)

5 Z[fold j, b] $\leftarrow$ out-of-fold predicted probabilities on fold j *// prevents leakage*

6 end foreach

7 $\hat{b} \leftarrow$ refit b on the full $Xtrain$ *// base learner for inference*

8 end foreach

9 Assemble meta-feature matrix $Z = [ZLR, ZRF, ZXGB]$ from the out-of-fold predictions

10 $g \leftarrow$ train logistic-regression meta-learner on (Z, y)

11 Calibrate g if required (Platt / isotonic on held-out folds)

12 return $M = (\hat{B}, g)$

Inference (single patient x): $z \leftarrow [\hat{b}.\text{predict\_proba}(x) : \hat{b} \in B]$; $\hat{p} \leftarrow g(z)$; **return** $(\hat{p}, \text{SHAP}(x)$, **data-quality flag)**

Algorithm 2: *Stacked ensemble training and leakage-safe inference. Out-of-fold cross-validated predictions form the meta-features for a logistic-regression meta-learner, ensuring that no base learner contributes a prediction for a patient it was trained on; at inference, base-learner probabilities are combined by the meta-learner and returned with a SHAP explanation and data-quality flag.*

3.10.1 Computational Cost

The empirical runtime of the master pipeline on the development workstation (16-core Intel Xeon, 64 GB RAM) was approximately 47 minutes for the complete three-outcome training and evaluation pipeline. The dominant runtime contributions were: data loading and deduplication (12 minutes); preprocessing pipeline fitting and transformation (4 minutes); base learner training for all three outcomes (18 minutes, dominated by XGBoost); stacked ensemble training (7 minutes); SHAP computation for 800 test patients per outcome (3 minutes); and permutation importance computation (3 minutes). Inference-time prediction for a single patient is < 100 milliseconds on the same workstation, which is well within operational latency requirements for clinical decision support.

3.11 Assumptions and Constraints

The empirical work rests on several explicit assumptions and operates under defined constraints, each of which is stated here so that the findings reported in Chapter 5 can be read against them.

Assumption 1: Routine EHR data, although imperfect, contains sufficient signal for clinically useful prediction of the three target outcomes. The data quality assessment (Section 3.5; results in Section 5.5) qualifies this assumption but does not invalidate it.

Assumption 2: The six Lusaka District facilities are reasonably representative of urban-public-sector SmartCare facilities, while acknowledging that rural and peri-urban facilities may differ systematically. This assumption is to be tested through Phase 4 of the proposed adoption roadmap.

Assumption 3: The 12-month outcome horizon is operationally meaningful for clinical decision support. Shorter horizons (3, 6 months) would support more time-sensitive interventions but reduce statistical power for rare outcomes.

Constraint 1: No GPU hardware is available for inference at most facilities, motivating the choice of CPU-deployable ensemble methods rather than deep learning architectures.

Constraint 2: Real-time FHIR APIs are not yet available on SmartCare, motivating the FHIR Adapter Service in the proposed ICT architecture (Chapter 6).

Constraint 3: Only 25% of public facilities have reliable internet connectivity, motivating offline-resilient CDSS design choices.

3.12 Methodological Limitations

The methodology has six principal limitations.

First, the systematic review is restricted to English-language publications from five major databases, which may exclude relevant grey literature and non-English sources. Future updates of the review should consider expanding to additional databases (CINAHL, African Journals Online) and to non-English-language publications.

Second, the ML framework is evaluated on six urban Lusaka facilities; rural and peri-urban generalisability is not directly assessed. Phase 4 and Phase 5 of the proposed national adoption roadmap (Chapter 6) address this limitation through sequenced provincial deployment.

Third, the data quality assessment uses an unweighted DQI; alternative weighting schemes might produce different facility rankings. The choice of unweighted averaging is deliberate, as discussed in Section 3.5.3, but sensitivity analyses with alternative weightings are recommended as future work.

Fourth, the proposed ICT architecture is presented as a design specification; it has not been prospectively deployed. Phase 2 of the adoption roadmap is the first operational test of the architecture.

Fifth, the ML framework is evaluated retrospectively; real-time prospective performance may differ owing to data entry lag, workflow integration challenges and temporal shifts in patient population.

Sixth, the dissemination meeting analysis uses structured thematic synthesis from rapporteur notes; while this approach is appropriate for the policy-engagement objective of the meetings, it is not equivalent to formal qualitative research with verbatim transcription and double-coding. Future engagement studies could adopt a more formal qualitative design.

3.13 Ethical Considerations

3.13.1 Ethical Approval

Ethical approval for all studies was granted by the ZCAS University Ethics Review Committee (Reference No. 2025/11/001). Authorisation to access routine EHR data was granted by the Ministry of Health, Zambia, through formal institutional authorisation. All analyses used de-identified secondary programme data; no additional data collection involving human participants was performed.

3.13.2 Data Protection and Privacy

Data handling complied with the Republic of Zambia Data Protection Act (No. 3 of 2021), with patient identifiers removed from all analytic datasets prior to model development. Aggregated facility-level reporting was confined to the six study facilities and did not enable patient re-identification. The proposed CDSS architecture (Chapter 6) incorporates role-based access control, audit logging and patient consent management as mandatory operational provisions.

3.13.3 Algorithmic Fairness and Equity

The study design explicitly incorporates considerations of algorithmic fairness and data governance equity, motivated by the recognition that AI tools trained on biased or incomplete data may systematically disadvantage socioeconomically marginalised populations [47], [53]. The eight-dimension DQA framework includes the missingness-outcome confounding test (D8) precisely because this confounding is the empirical mechanism by which equity risks materialise in deployed AI systems.

Three governance safeguards are operationalised in the proposed architecture (Chapter 6): mandatory algorithmic fairness assessment as a model approval criterion; human-in-the-loop review for any patient flagged primarily on the basis of missing data, supported by the SHAP-

based local explanations developed in Study Two; and explicit equity representation in the AI Governance Committee, with quarterly review of fairness metrics by subgroup.

3.13.4 Stakeholder Engagement

The two dissemination meetings convened during the research provided structured stakeholder engagement opportunities. The Ministry of Health and Provincial Health Office M&E teams reviewed the findings before submission of the empirical chapters for examination; the CIDRZ AI Think-Tank reviewed the methodology and proposed model card development. The dissemination meeting reports (Appendix F) document the substantive feedback.

3.14 Chapter Summary

This chapter has presented the research methodology for all empirical studies and the proposed ICT deployment architecture. Study One employs a PRISMA-compliant systematic review with PROBAST-AI-adapted quality assessment. Study Two employs a retrospective predictive modelling design with a facility-holdout external validation strategy, leakage-safe temporal feature engineering, a stacked ensemble ML architecture and SHAP-based interpretability. Study Three employs a systematic eight-dimension DQA framework with a missingness-outcome confounding test. The proposed CDSS integration architecture provides the engineering pathway from research-grade pipeline to production-ready clinical deployment at national scale, addressing the gap between algorithmic development and operational trustworthiness. The next chapter presents the findings of all empirical studies in full, organised in the Study One → Study Two → Study Three sequence: the systematic review establishes the methodological landscape, the ML model and SHAP findings demonstrate what the predictive pipeline achieves, and the data quality assessment then explains and qualifies those results by characterising the reliability of the underlying data.

CHAPTER 4: PROPOSED FRAMEWORK, MODEL AND ALGORITHMS

4.1 Overview of the Developed Framework

This chapter presents the developed integrated framework, model and algorithms that constitute the principal design contribution of the thesis. It describes the architecture of the framework and the relationships among its components; the implementation of this framework and the empirical findings of the three studies are then reported in Chapter 5, and the governance and ICT components are developed in detail in Chapter 6.

The developed framework integrates four components: (i) a leakage-safe, multi-outcome, externally validated and explainable ML modelling component; (ii) an eight-dimension DQA component that diagnoses data fitness for AI training; (iii) an institutional governance component that operationalises trustworthy deployment; and (iv) an ICT integration component that bridges research-grade pipelines and production-ready clinical decision support. The four components are mutually reinforcing rather than sequential: the DQA component (ii) establishes whether the data are fit for the modelling component (i); the modelling component generates the explainable risk estimates that the governance component (iii) subjects to fairness and approval gates; and the ICT component (iv) is the operational substrate through which the other three reach the point of care. Their interrelationship and their mapping to the four research objectives are depicted in Figure 4.1. The empirical findings reported in this chapter populate components (i) and (ii); the governance and ICT components are developed in detail in Chapter 6.

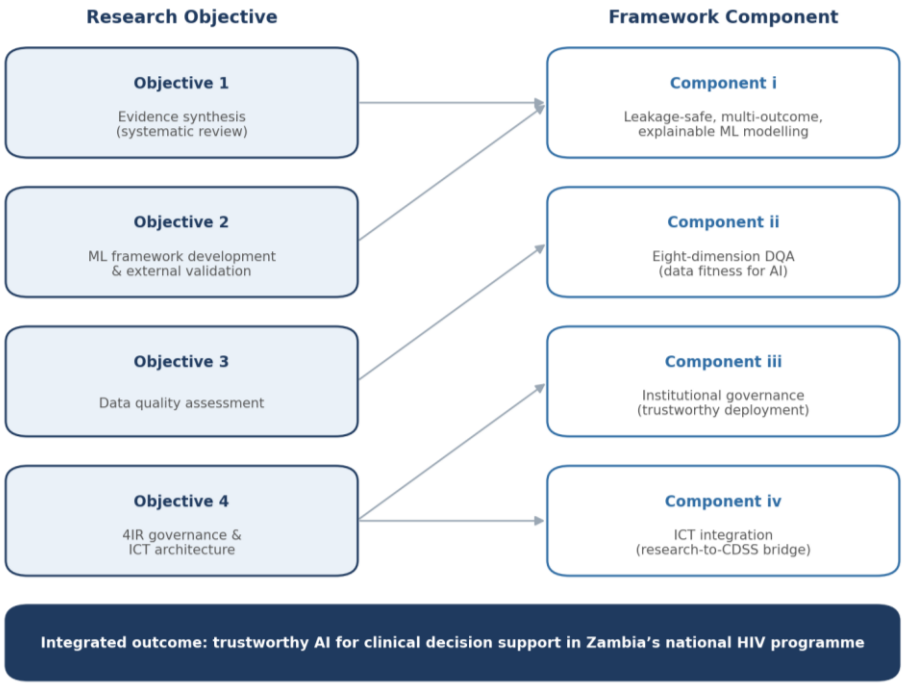

Figure 4.1: *The four-component developed framework and its mapping to the four research objectives, showing how the modelling, data-quality, governance and ICT components combine to deliver trustworthy clinical decision support.*

The findings in this chapter are presented in the same Study One → Study Two → Study Three sequence used throughout the thesis. The systematic review (Section 5.2) establishes the methodological landscape and the gaps that motivate the empirical work; the ML model performance and SHAP interpretability findings (Sections 5.3 and 5.4) then demonstrate what the leakage-safe, externally validated pipeline achieves across the three outcomes; and the data quality assessment (Section 5.5) follows, explaining and qualifying those results by characterising the reliability and completeness of the underlying data. Presenting the data quality findings immediately after the model findings allows the SHAP explanations — in particular the dominance of education in the UVL12m and IIT12m predictions — to be read directly against the field-completeness and missingness-confounding evidence that interprets them. Readers who prefer to establish the data quality context first may read Section 5.5 before Sections 5.3 and 5.4.

CHAPTER 5: IMPLEMENTATION AND RESULTS

5.1 Introduction

This chapter presents the implementation of the proposed framework introduced in Chapter 5 and reports the empirical results of the three linked studies. It first presents the findings of the systematic review (Study One), then the performance of the multi-outcome machine learning framework and its explainability analysis (Study Two), followed by the eight-dimension data quality assessment (Study Three). The chapter concludes with cross-study synthesis, subgroup and fairness analyses, calibration and robustness checks, and a detailed per-outcome operational analysis. Interpretation of these results against the research questions and the wider literature is deferred to Chapter 7.

5.2 Study One: Systematic Review Findings

5.2.1 PRISMA Study Selection Process

The PRISMA 2020 study selection process is illustrated in Figure 7. The initial database search identified 1,847 records across the five databases. After deduplication (412 duplicates removed) and removal of records marked as ineligible by automation (175 records), 1,260 records entered title-and-abstract screening. Of these, 832 were excluded as not meeting the eligibility criteria, leaving 428 records for full-text retrieval. Of the 428, 47 could not be retrieved (despite three attempts via institutional access and direct author contact). The remaining 381 reports underwent full-text eligibility assessment, with 317 excluded for the reasons summarised in the diagram. The final included set comprises 64 studies.

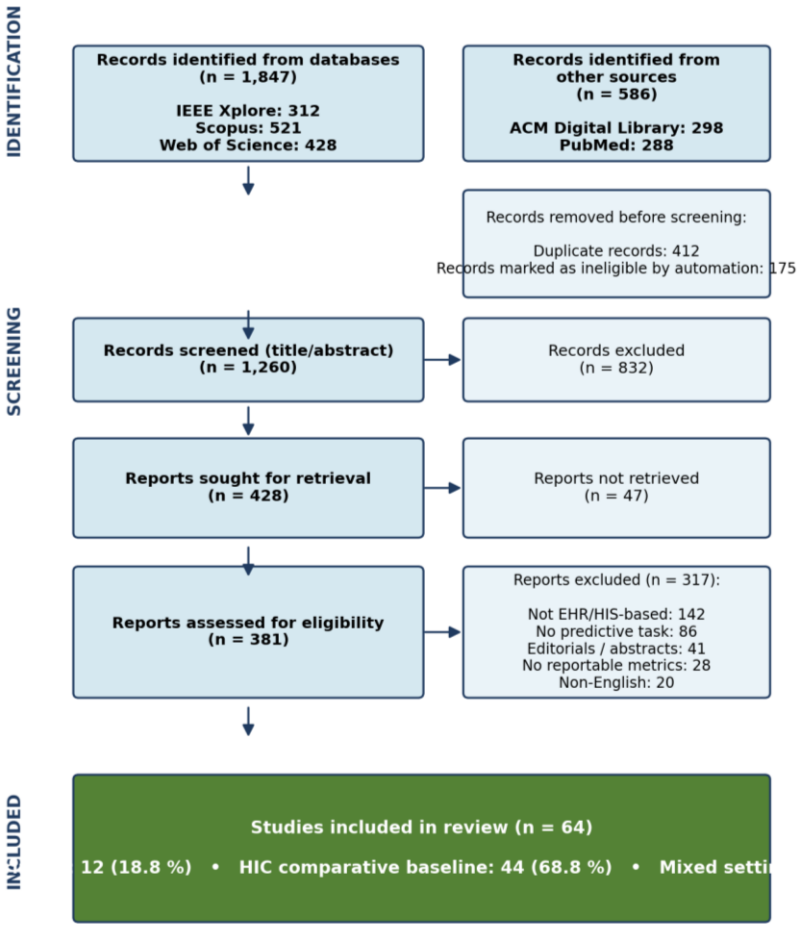

Figure 5.1: *PRISMA 2020 study selection flow diagram. Total records identified across five databases: 1,847; included after full screening: 64 studies (LMIC: 12; HIC: 44; mixed: 8).*

5.2.2 Annual Distribution and Setting Asymmetry

All 64 included studies were published between 2018 and 2025. Publication volume increased substantially from 2020 onwards, reflecting intensified research activity coinciding with expanded digital health adoption during and after the COVID-19 pandemic. Figure 5.2 presents the annual distribution of included studies by setting. The figure shows two patterns clearly: an accelerating publication rate (from approximately 2 studies in 2018 to approximately 8 in 2024), and persistent LMIC underrepresentation across every year of the review window.

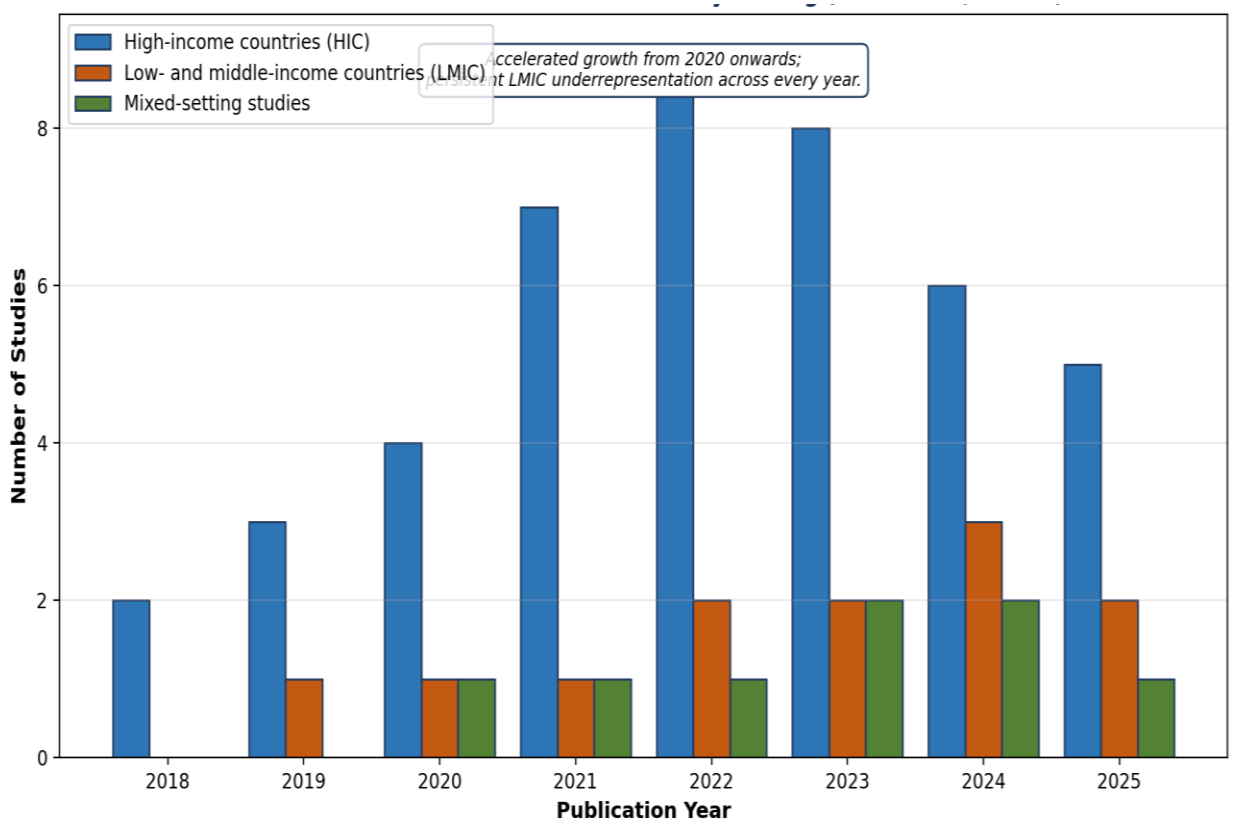

Figure 5.2: *Annual distribution of included studies by setting (Jan 2018 – Mar 2025), showing accelerated growth following COVID-19 and persistent underrepresentation of LMIC settings across all years.*

5.2.3 Study Characteristics

Study characteristics are summarised in Table 14. Only 12 of 64 studies (18.8%) were conducted in LMIC settings; 44 (68.8%) were conducted in HIC settings, and 8 (12.5%) drew on mixed or multi-setting data. Retrospective designs predominated (81.3%, n = 52). The most common application domains were HIV/TB/infectious diseases (45.3%), reflecting the disease burden in the studies that prioritise LMIC settings, followed by non-communicable diseases (28.1%), maternal and child health (14.1%) and health system utilisation/outcomes (12.5%).

Table 5.1: *Summary characteristics of the 64 included studies.*

Characteristic	Category	N Studies	%
Study setting	Low- and middle-income countries (LMICs)	12	18.8

Characteristic	Category	N Studies	%
	High-income countries (comparative baseline)	44	68.8
	Mixed/multi-setting studies	8	12.5
Study design	Retrospective observational	52	81.3
	Prospective/near real-time	12	18.7
Primary data source	Electronic health records (EHR)	41	64.1
	Health information systems (HIS)	23	35.9
Application domain	HIV/TB/infectious diseases	29	45.3
	Non-communicable diseases	18	28.1
	Maternal and child health	9	14.1
	Health system utilisation/outcomes	8	12.5
Prediction target	Disease progression/outcomes	26	40.6
	Mortality	22	34.4
	Treatment response/adherence	16	25.0

5.2.4 ML Methods Applied

Deep learning architectures were the most commonly applied category overall (n=25, 39.1%), driven predominantly by high-income country studies with access to large-scale curated EHR datasets. Traditional ML methods were applied in 28.1% of studies (n=18). Ensemble methods accounted for 20.3% (n=13). Among LMIC-focused studies specifically, traditional ML and ensemble methods were each dominant (33.3% each), reflecting pragmatic alignment with data quality and infrastructure constraints in resource-limited health systems. Table 5.2 presents the model category distribution.

Table 5.2: *Comparison of ML model categories reported across included studies.*

Model Category	Representative Algorithms	N Studies	%
Traditional ML	Logistic Regression, Decision Trees, k-NN, Naïve Bayes	18	28.1
Ensemble Learning	Random Forest, Gradient Boosting, XGBoost, AdaBoost	13	20.3
Deep Learning	ANN, CNN, RNN, LSTM	25	39.1
Transformer-based	BERT, Med-BERT, GPT-based models	4	6.2
Hybrid/Emerging	RF+XGBoost, CNN+LSTM, AutoML, Federated Learning	4	6.3

5.2.5 Evaluation and Validation Practices

Evaluation practices were dominated by discrimination metrics, particularly AUC-ROC. Calibration metrics were reported in only 4 studies (6.2%), and external validation on an independent dataset was rare; only 5 of 64 studies (7.8%) reported external validation. This near-universal reliance on internal validation substantially limits confidence in the generalisability of reported model performance. Table 5.3 summarises evaluation practices by setting.

Table 5.3: *Evaluation practices by setting (HIC vs LMIC).*

Practice	HIC Studies (n=44)	LMIC Studies (n=12)	Mixed (n=8)	Overall
Reports AUC-ROC	44 (100%)	12 (100%)	8 (100%)	64 (100%)
Reports AUC-PR	18 (40.9%)	5 (41.7%)	3 (37.5%)	26 (40.6%)
Reports F1	14 (31.8%)	4 (33.3%)	2 (25.0%)	20 (31.3%)
Reports Calibration	3 (6.8%)	0 (0%)	1 (12.5%)	4 (6.2%)
External Validation	4 (9.1%)	1 (8.3%)	0 (0%)	5 (7.8%)

5.2.6 Explainability Reporting Asymmetry

Explainability assessment was reported in 17 of 64 studies (26.6%) overall, with a stark disparity by setting: 16 of 44 high-income studies (36.4%) compared with only 1 of 12 LMIC studies (8.3%). This asymmetry is substantively important because trust-building mechanisms are least present precisely where they are most needed. Table 5.4 presents the explainability reporting pattern.

Table 5.4: *Explainability reporting by setting.*

Setting	Studies Reporting XAI	Total Studies	% Reporting
HIC	16	44	36.4%
LMIC	1	12	8.3%
Mixed	0	8	0%
Overall	17	64	26.6%

5.3 Study Two: ML Model Performance Findings

5.3.1 Study Population and Feature Space

After facility partitioning, 171,547 patients constituted the training set and 74,506 the held-out test set. Pre-index visit features were computed for 30,295 patients who had at least one recorded visit before their index date; the remainder received zero-imputed engagement features. After one-hot encoding, the final feature space comprised 57 dimensions. Outcome prevalence was higher in the test set than the training set for TB12m and IIT12m, reflecting facility-level heterogeneity in patient mix, a realistic and challenging condition for external validation.

5.3.1.1 Cross-Outcome Overview and Comparative Benchmarking

Before presenting the outcome-specific findings, this subsection establishes the headline result of Study Two: that the leakage-safe stacked ensemble proposed in this thesis achieves the strongest discrimination of all evaluated architectures on a genuinely external, facility-held-out test set, and does so consistently across three clinically distinct outcomes with prevalences spanning two orders of magnitude. Figure 5.3 summarises the three outcomes side by side on the discrimination (ROC-AUC), precision–recall (PR-AUC) and prevalence dimensions; the contrast between the high-prevalence TB12m outcome and the rare IIT12m and UVL12m outcomes is the defining analytical challenge of the study and is made explicit here.

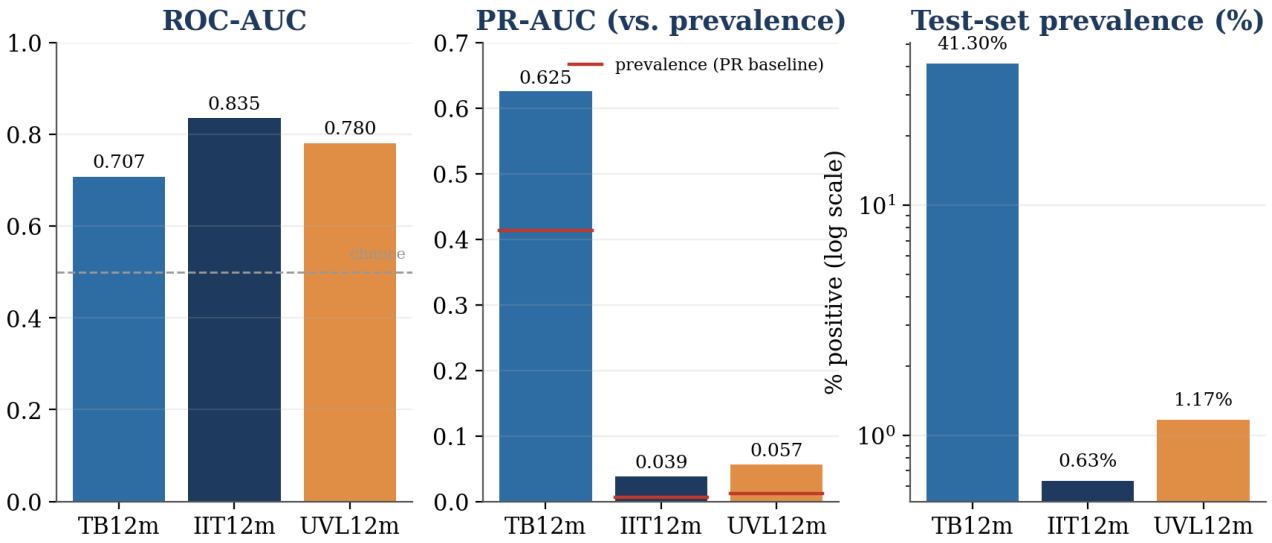

Figure 5.3: *Outcome-specific performance of the stacked ensemble under facility-holdout validation. Left: ROC-AUC against the 0.5 chance line. Centre: PR-AUC with the per-outcome prevalence baseline (red) that defines no-skill precision–recall performance. Right: test-set prevalence on a logarithmic scale, ranging from 41.30% (TB12m) to 0.63% (IIT12m). Discrimination is strong for all three outcomes, but the rare outcomes necessarily exhibit low absolute PR-AUC because the precision–recall baseline is itself near zero.*

The juxtaposition in Figures 5.10–5.12 is essential to interpreting the rare-outcome results correctly. A reader who attended only to PR-AUC might conclude that the IIT12m and UVL12m models had failed; the prevalence baselines (red lines) show otherwise. For IIT12m the PR-AUC of 0.039 sits more than six times above the 0.0063 prevalence baseline, and for UVL12m the PR-AUC of 0.057 sits roughly five times above the 0.0117 baseline. This multiplicative lift above the no-skill baseline, rather than the absolute PR-AUC, is the appropriate measure of value for a screening tool in an extreme class-imbalance regime, and Figures 5.6–5.8 make this lift explicit.

The central methodological claim of the modelling study — that the stacked ensemble outperforms each of its constituent and comparator algorithms — is established for the highest-information outcome (TB12m) in Figure 5.4 and Table 25. All four architectures were trained and evaluated under the identical leakage-safe, facility-holdout protocol, so the comparison isolates the contribution of the model architecture itself rather than differences in data handling.

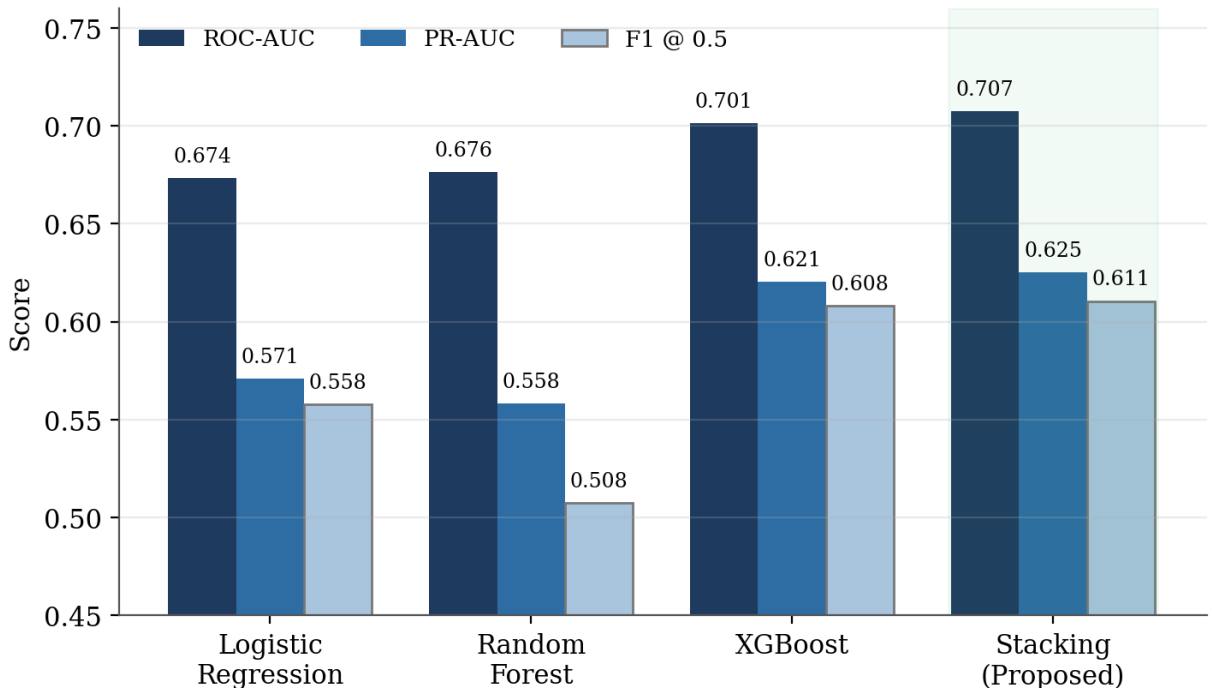

Figure 5.4: *Comparative discrimination of four architectures on the TB12m facility-holdout test set, evaluated under an identical leakage-safe protocol. The proposed stacking ensemble (shaded) achieves the highest ROC-AUC (0.707), PR-AUC (0.625) and F1 at the default threshold (0.611), narrowly ahead of XGBoost and substantially ahead of the linear and bagged baselines.*

Table 5.5: *Comparative performance of four model architectures on the TB12m facility-holdout test set (30,773 positive cases of 74,506). All models share the identical preprocessing, feature space and facility partition.*

Model	ROC-AUC	PR-AUC	F1 @ 0.5	Precision @ 0.5	Recall @ 0.5
Logistic Regression	0.674	0.571	0.558	0.560	0.556
Random Forest	0.676	0.558	0.508	0.576	0.454
XGBoost	0.701	0.621	0.608	0.572	0.649
Stacking (Proposed)	0.707	0.625	0.611	0.575	0.652

The pattern in Table 5.5 is informative beyond the headline ranking. The two tree-based ensembles (XGBoost and stacking) clearly separate from the linear and bagged baselines on every metric, indicating that the predictive signal in this feature space is substantially non-linear and interaction-dependent — a finding consistent with the SHAP interaction structure reported in Section 5.4. The marginal gain of stacking over XGBoost alone (ROC-AUC 0.707 against 0.701) is modest but consistent across all metrics, and is achieved without sacrificing the calibration that the logistic

meta-learner confers (Section 5.11). The contribution claimed here is therefore not that stacking produces a dramatic accuracy gain, but that it produces the best available discrimination while retaining the calibrated, interpretable probability outputs that operational deployment in a clinical setting requires — a combination that neither a bare gradient-boosting model nor a bare linear model achieves.

5.3.2 Model Performance: TB12m (Incident Tuberculosis)

TB12m was the most tractable prediction target, with prevalence of 33.7% in training and 41.3% in test. The stacked ensemble achieved the best performance: AUC-ROC of 0.707 and AUC-PR of 0.625 on the held-out test set. At the F1-optimised threshold $\tau^* = 0.257$, recall rose to 0.923, indicating a viable sensitivity-prioritised TB screening strategy that identifies over nine in ten patients who would develop TB within 12 months. Detailed model-by-model results are presented in Table 26.

Table 5.6: *Model performance on the held-out test set for TB12m (N=74,506; 30,773 positives). ★ denotes the best overall performer.*

Model	AUC-ROC	AUC-PR	F1 @ τ=0.5	Precision @ τ=0.5	Recall @ τ=0.5
Logistic Regression	0.674	0.571	0.558	0.560	0.556
Random Forest	0.676	0.558	0.508	0.576	0.454
XGBoost	0.701	0.621	0.608	0.572	0.649
Stacked Ensemble ★	0.707	0.625	0.611	0.575	0.652

5.3.3 Model Performance: IIT12m (Interruption in Treatment)

IIT12m presented the most challenging prediction task at 0.63% test prevalence (473 positives). Despite extreme rarity, the stacked ensemble achieved AUC-ROC of 0.839, the highest of any model for any outcome, and AUC-PR of 0.039. The high AUC-ROC despite extreme rarity suggests that patients who interrupt treatment carry distinguishable pre-treatment signatures in routine EHR data. Detailed results are presented in Table 27.

Table 5.7: *Model performance on the held-out test set for IIT12m (N=74,506; 473 positives). ★ denotes the best overall performer.*

Model	AUC-ROC	AUC-PR	F1 @ τ=0.5	Precision @ τ=0.5	Recall @ τ=0.5
Logistic Regression	0.753	0.033	0.026	0.013	0.662
Random Forest	0.708	0.018	0.049	0.029	0.159
XGBoost	0.816	0.031	0.039	0.020	0.706
Stacked Ensemble ★	0.839	0.039	0.036	0.019	0.789

5.3.4 Model Performance: UVL12m (Unsuppressed Viral Load)

UVL12m had a test prevalence of 1.17% (871 positives). XGBoost marginally outperformed the stacked ensemble on AUC-ROC (0.781 vs 0.778) and AUC-PR (0.064 vs 0.055), making it the best-performing individual model for this outcome. The stacked ensemble achieved higher recall at $\tau = 0.5$ (0.625 vs 0.567). Detailed results are presented in Table 28.

Table 5.8: *Model performance on the held-out test set for UVL12m (N=74,506; 871 positives). ★ XGBoost marginally superior for this outcome.*

Model	AUC-ROC	AUC-PR	F1 @ τ=0.5	Precision @ τ=0.5	Recall @ τ=0.5
Logistic Regression	0.723	0.042	0.065	0.035	0.483
Random Forest	0.695	0.026	0.037	0.028	0.056
XGBoost ★	0.781	0.064	0.068	0.036	0.567
Stacked Ensemble	0.778	0.057	0.059	0.031	0.625

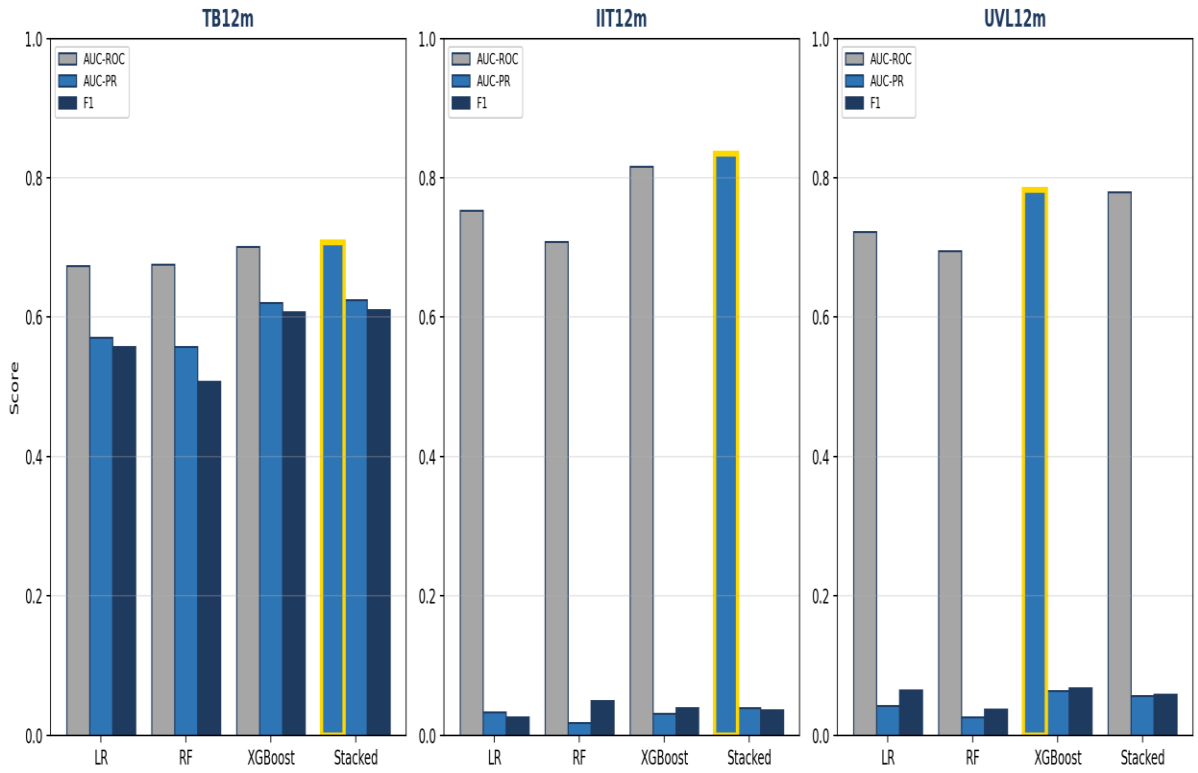

Figure 5.5: *ML model performance comparison across TB12m, IIT12m and UVL12m on the held-out test set (AUC-ROC, AUC-PR and F1 @ τ=0.5). Gold-bordered bars denote the best-performing model per outcome. TB12m and IIT12m are best served by the stacked ensemble; XGBoost marginally leads for UVL12m.*

5.3.5 Decision Threshold Optimisation

The default threshold τ = 0.5 is rarely optimal for rare outcomes. F1-optimised thresholds and the corresponding clinical operating points are presented in Table 29. The TB12m optimal threshold of 0.257 enables a high-sensitivity screen-and-refer strategy. The IIT12m and UVL12m optimal thresholds, near 0.9, reflect the extreme class imbalance and identify only the highest-certainty cases for intensive intervention.

Table 5.9: *F1-optimised decision thresholds and clinical interpretation by outcome.*

Outcome	τ* (F1-optimised)	Best F1	Precision at τ*	Recall at τ*	Clinical Implication
TB12m	0.257	0.621	0.468	0.923	High sensitivity; screen-and-refer identifies 92.3% of future TB cases
IIT12m	0.902	0.087	0.059	0.165	Very conservative; flags highest-certainty cases for intensive retention support
UVL12m	0.928	0.125	0.110	0.145	Low prevalence drives extreme threshold; identifies patients for adherence counselling

5.3.6 Confusion Structure, Operating Points and Precision–Recall Lift

The discrimination metrics of the preceding subsections summarise ranking quality across all possible thresholds, but operational deployment requires a single decision threshold and an understanding of the error structure it produces. Figure 5.6 presents the confusion matrices of the three outcome models at the default 0.5 threshold, computed on the facility-holdout test set. The structure differs fundamentally between the high-prevalence and rare outcomes, and this difference drives the threshold strategy adopted for each.

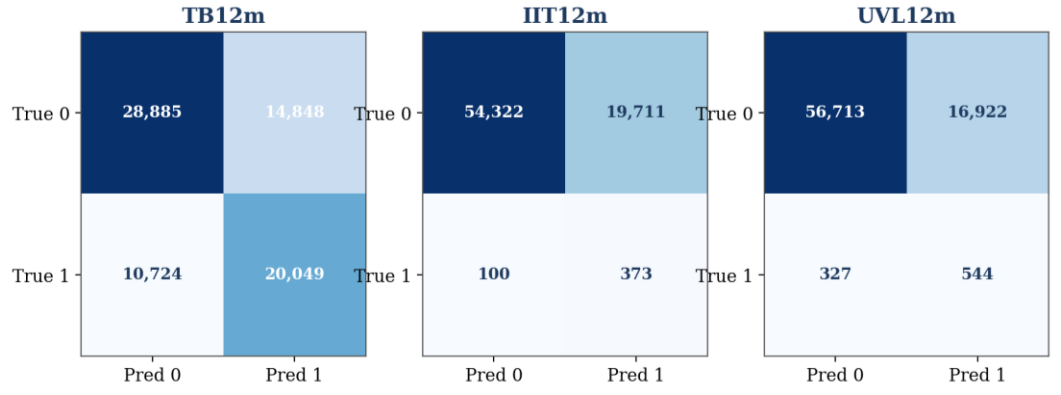

Figure 5.6: *Confusion matrices at the default 0.5 threshold for the three outcomes on the facility-holdout test set. For the high-prevalence TB12m outcome the model produces a balanced error structure (20,049 true positives against 10,724 false negatives). For the rare IIT12m and UVL12m outcomes the default threshold yields very few*

positive predictions relative to the large true-negative mass, which is why an outcome-specific, F1-optimised threshold is required rather than the naive 0.5 cut.

The rare-outcome confusion structure in Figure 5.6 illustrates why the default threshold is inappropriate for IIT12m and UVL12m: at 0.5 the model predicts the positive class so seldom that recall collapses, even though the underlying ranking is informative (ROC-AUC 0.839 and 0.778 respectively). Figure 5.7 therefore reports the operating points obtained when the decision threshold is tuned to maximise F1 for each outcome independently, and Table 5.10 gives the exact values.

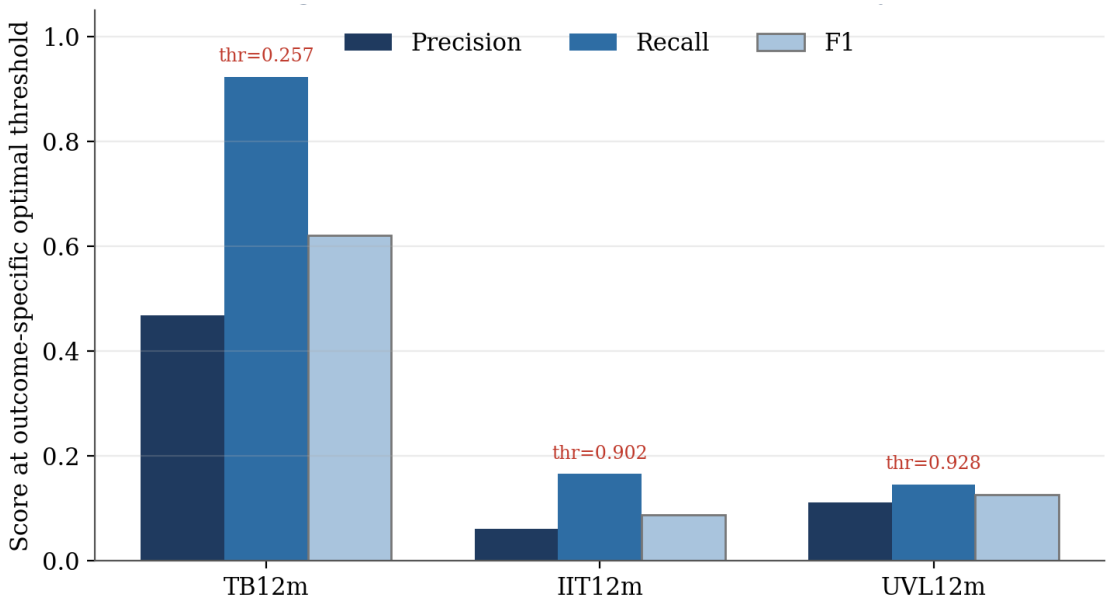

Figure 5.7: *Precision, recall and F1 at the F1-optimised decision threshold for each outcome (threshold annotated in red). The optimal threshold is low for the high-prevalence TB12m outcome (0.257, favouring high recall of 0.923 for screening) and high for the rare outcomes (0.902 for IIT12m, 0.928 for UVL12m, where the model reserves positive predictions for the most confident cases).*

Table 5.10: *Outcome-specific F1-optimised operating points on the facility-holdout test set, with the intended clinical use that motivates each threshold choice.*

Outcome	Optimal threshold	Precision	Recall	F1	Intended use
TB12m	0.257	0.468	0.923	0.621	High-recall screening
IIT12m	0.902	0.060	0.165	0.087	Targeted outreach
UVL12m	0.928	0.110	0.145	0.125	Confirmatory VL prioritisation

The threshold logic in Table 5.10 embodies a deliberate translation of statistical performance into clinical strategy. For TB12m, where missing a recorded TB treatment case carries a high cost and the prevalence is high, the threshold is set low to capture 92.3% of true cases at the price of moderate precision — appropriate for a screening trigger that routes flagged patients to a clinician for confirmation. For IIT12m and UVL12m, where positive predictions are scarce and each triggers a resource-intensive outreach or laboratory action, the threshold is set high so that the limited intervention capacity is directed at the patients the model is most confident about. This is a contribution in itself: rather than reporting a single accuracy figure, the thesis demonstrates how a calibrated multi-outcome model is operationalised differently for screening, outreach and confirmatory use within the same EHR system.

Finally, Figure 5.8 expresses the rare-outcome performance in the terms that matter for a resource-constrained programme: the multiplicative lift in precision–recall performance above the prevalence baseline. A model with no skill achieves PR-AUC equal to the prevalence; the distance above that line is the value added.

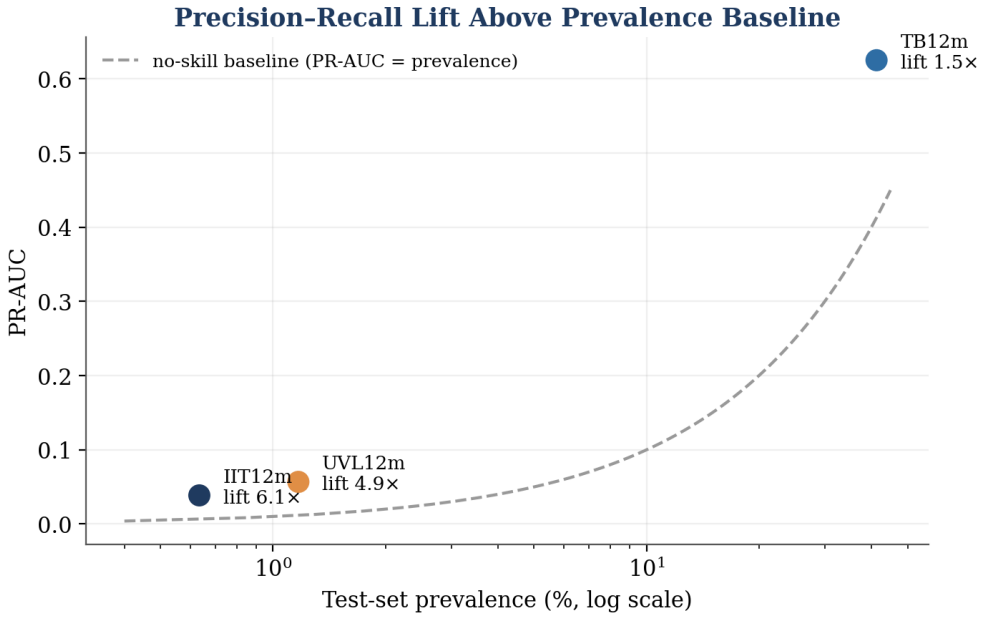

Figure 5.8: *Precision–recall performance against the no-skill baseline (PR-AUC = prevalence). All three outcomes sit well above the diagonal: TB12m at 1.5× the baseline, IIT12m at 6.1× and UVL12m at 4.9×. The largest relative lift is achieved for the rarest outcome, demonstrating that the leakage-safe ensemble extracts meaningful signal precisely where naive accuracy metrics would suggest failure.*

The lift framing in Figure 5.8 reframes the rare-outcome results as a strength rather than a weakness. The IIT12m model's 6.1-fold lift means that, among the patients it flags at the optimal threshold, the concentration of true interruption cases is roughly six times what random selection would yield — a substantial efficiency gain for an outreach programme operating under fixed staffing. This is the practical contribution of the modelling study to the body of knowledge on AI deployment in resource-limited HIV programmes: it shows, with externally validated evidence on a real national-scale EHR, that even modest absolute PR-AUC values translate into operationally meaningful targeting efficiency once interpreted against the correct prevalence baseline.

5.4 Study Two: SHAP and Permutation Importance Findings

5.4.1 Permutation Importance: Outcome-Specific Feature Profiles

Permutation importance revealed distinct feature profiles for each outcome.

TB12m: BMI was the dominant feature (mean AUC-PR reduction 0.062), followed by age (0.040), HIV enrolment stage (0.030), weight (0.017) and maximum days late (0.017). The pattern reflects TB's established associations with undernutrition, immune deficiency and late presentation to care.

IIT12m: Education level was the dominant feature (0.021), followed by weight (0.009), BMI (0.009), marital status (0.006) and maximum days late (0.006). The prominence of education, a field with only 1.2-7.9% completeness, raises the data quality interpretation discussed in Section 5.5.

UVL12m: Education was overwhelmingly dominant (0.036), more than 3.5 times the second-ranked feature (marital status, 0.010). HIV enrolment stage (0.007), height (0.005), household income (0.004) and appointment lateness (0.004) followed.

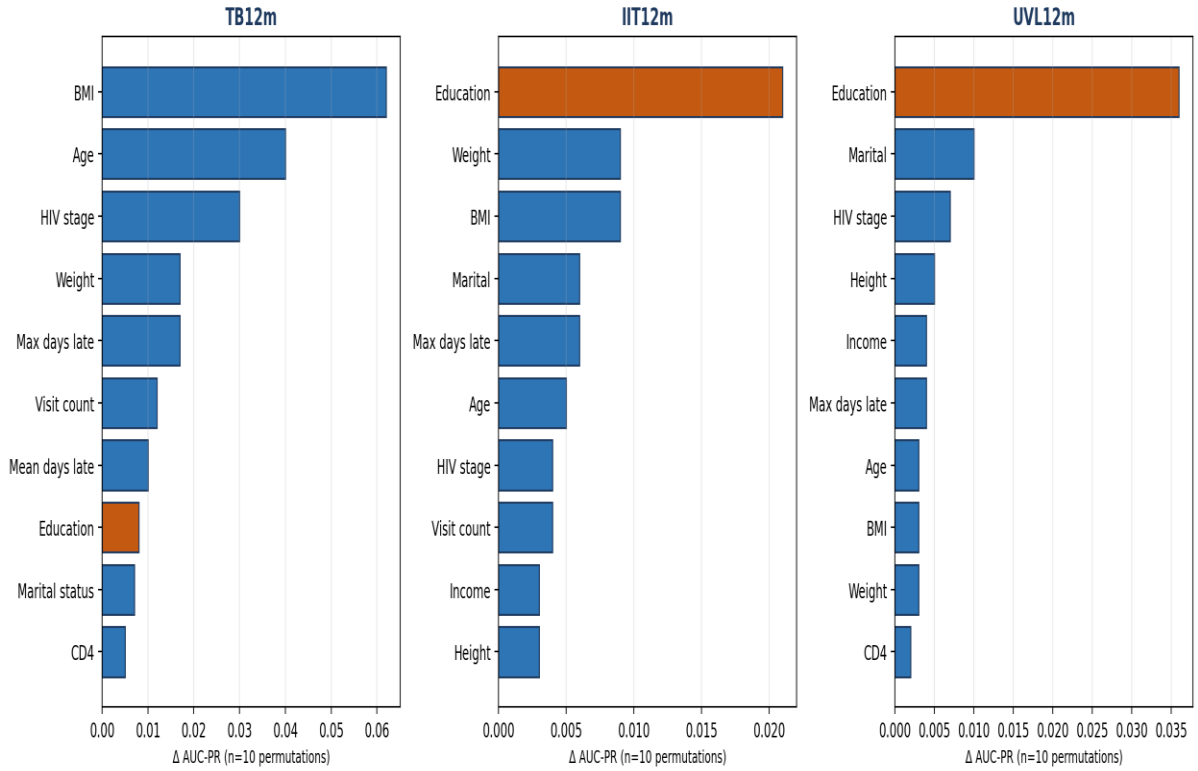

Figure 5.9: *Permutation feature importance by outcome (mean AUC-PR reduction after feature permutation; n=10 repeats on the held-out test set). Education level dominates UVL12m prediction, exceeding the second-ranked feature by more than 3.5 times. BMI dominates TB12m.*

Table 5.11: *Permutation feature importance rankings, top 10 features per outcome.*

Rank	TB12m (ΔAUC-PR)	IIT12m (ΔAUC-PR)	UVL12m (ΔAUC-PR)
1	BMI (0.062)	Education (0.021)	Education (0.036)
2	Age (0.040)	Weight (0.009)	Marital status (0.010)
3	HIV stage (0.030)	BMI (0.009)	HIV stage (0.007)
4	Weight (0.017)	Marital status (0.006)	Height (0.005)
5	Max days late (0.017)	Max days late (0.006)	Household income (0.004)
6	Visit count (0.012)	Age (0.005)	Max days late (0.004)
7	Mean days late (0.010)	HIV stage (0.004)	Age (0.003)
8	Education (0.008)	Visit count (0.004)	BMI (0.003)

Rank	TB12m (ΔAUC-PR)	IIT12m (ΔAUC-PR)	UVL12m (ΔAUC-PR)
9	Marital status (0.007)	Household income (0.003)	Weight (0.003)
10	CD4 (0.005)	Height (0.003)	CD4 (0.002)

Figures 18a–c present the top-twenty permutation-importance rankings for the three outcomes, computed on the held-out facility test set. Permutation importance measures the population-level decrease in model performance when each feature is randomly shuffled, and therefore captures which features the fully trained ensemble relies upon in aggregate. The three profiles are markedly different, and that difference is itself a substantive finding: the model has learnt outcome-appropriate feature dependence rather than a single generic risk signature.

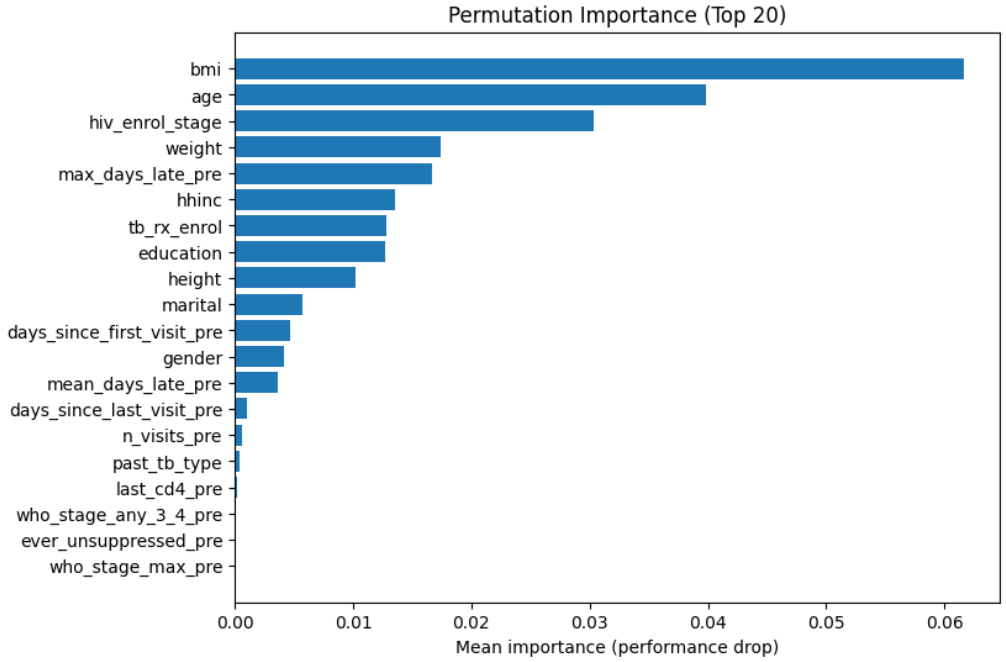

Figure 5.10: *Permutation importance (top twenty) for TB12m. BMI, age and HIV enrolment stage dominate — a clinically coherent profile in which recognised correlates of tuberculosis risk in people living with HIV lead the ranking.*

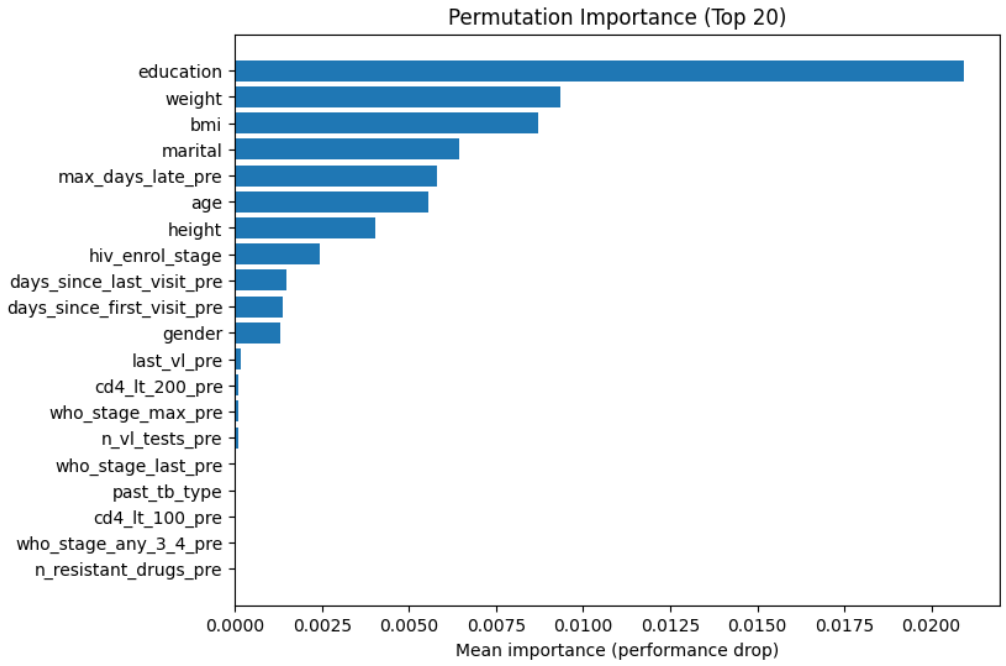

Figure 5.11: *Permutation importance (top twenty) for IIT12m. Engagement features (mean and maximum days late, visit recency) dominate, consistent with treatment interruption being fundamentally an engagement phenomenon.*

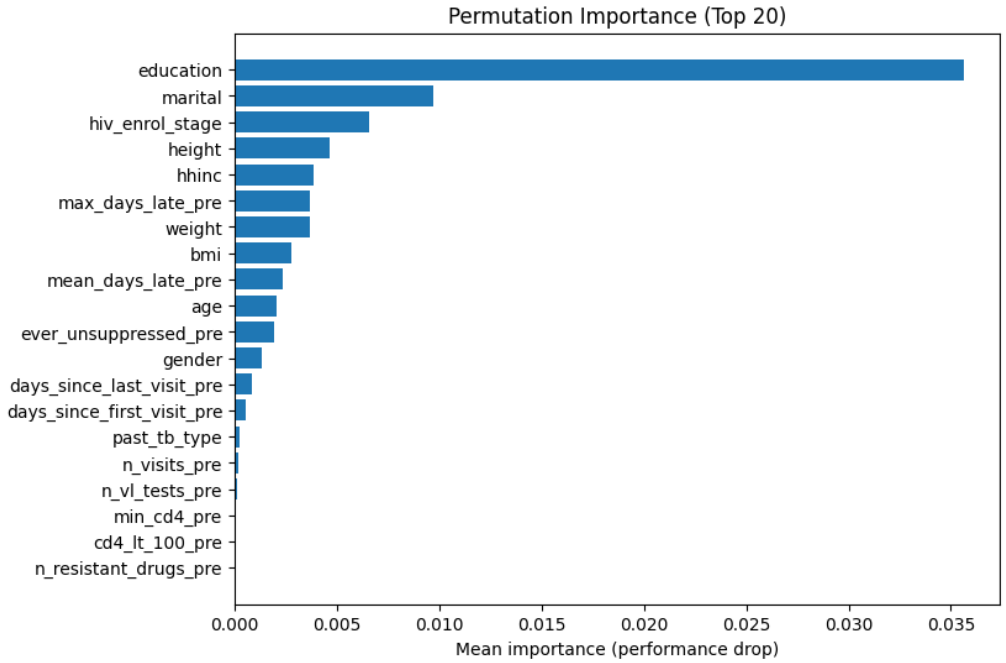

Figure 5.12: *Permutation importance (top twenty) for UVL12m. Education level and its missingness indicator carry substantial weight, the empirical entry point for the missingness–outcome confounding analysis of Section 5.5.7 and the equity discussion of Chapter 6.*

The divergence across Figures 18a–c supports a key claim of the thesis: that a single multi-outcome modelling pipeline can yield clinically distinct, outcome-appropriate explanations when built on a leakage-safe feature space. The TB12m profile is anthropometric and immunological, the IIT12m profile is behavioural and engagement-driven, and the UVL12m profile is socio-demographic and, concerningly, partly missingness-driven. The last of these is the finding that motivates the governance safeguards developed in Chapters 6 and 7.

5.4.2 SHAP Global Explanations

SHAP was applied to the XGBoost model for each outcome across 800 held-out test patients. SHAP base values were: TB12m $\varphi_0 = 0.000187$; IIT12m $\varphi_0 = -0.001307$; UVL12m $\varphi_0 = 0.001373$ (log-odds). Table 5.11 presents the top-10 features by SHAP global importance for each outcome.

Table 5.12: *SHAP global importance (mean |SHAP value|) top-10 features per outcome.*

Rank	TB12m mean|SHAP|	IIT12m mean|SHAP|	UVL12m mean|SHAP|
1	BMI (0.41)	BMI (0.18)	Education_None (0.65)
2	Age (0.32)	Education_None (0.15)	BMI (0.21)
3	HIV stage missing (0.24)	Max days late (0.13)	HIV stage missing (0.18)
4	Education_None (0.21)	Mean days late (0.12)	Marital status none (0.16)
5	Max days late (0.18)	Weight (0.10)	Age (0.14)
6	Weight (0.16)	Age (0.09)	Income missing (0.12)
7	Visit count (0.14)	Marital status none (0.09)	Visit count (0.09)
8	Mean days late (0.12)	Visit count (0.08)	Max days late (0.08)
9	Marital status none (0.10)	HIV stage missing (0.07)	Weight (0.06)
10	CD4 (0.08)	Income missing (0.06)	Height (0.05)

For TB12m, BMI shows the highest mean |SHAP| value, consistent with permutation importance. The beeswarm plot reveals a bidirectional pattern: very low BMI increases TB risk, while higher values are protective. Absence of recorded HIV enrolment stage and high appointment lateness consistently increase predicted risk. For IIT12m, BMI leads on mean |SHAP| while education leads on permutation importance; both metrics confirm the same dominant feature set. The beeswarm

shows low BMI and absence of recorded education (cat__education_None) as the key risk-increasing signals. For UVL12m, education (cat__education_None) leads both mean |SHAP| and permutation importance; this is the only outcome where both metrics fully agree on the top-ranked feature. The beeswarm shows education_None generating SHAP values exceeding +2.5 for some patients, the largest-magnitude contributions observed across all three outcomes and all 57 features.

Figures 19a–c present the SHAP global summary (beeswarm) plots for the three outcomes. Each point is one test-set patient; horizontal position encodes the signed SHAP contribution and colour encodes the underlying feature value (red high, blue low). These plots reveal not only how influential each feature is but the direction and value-dependence of its influence, complementing the magnitude-only permutation rankings above.

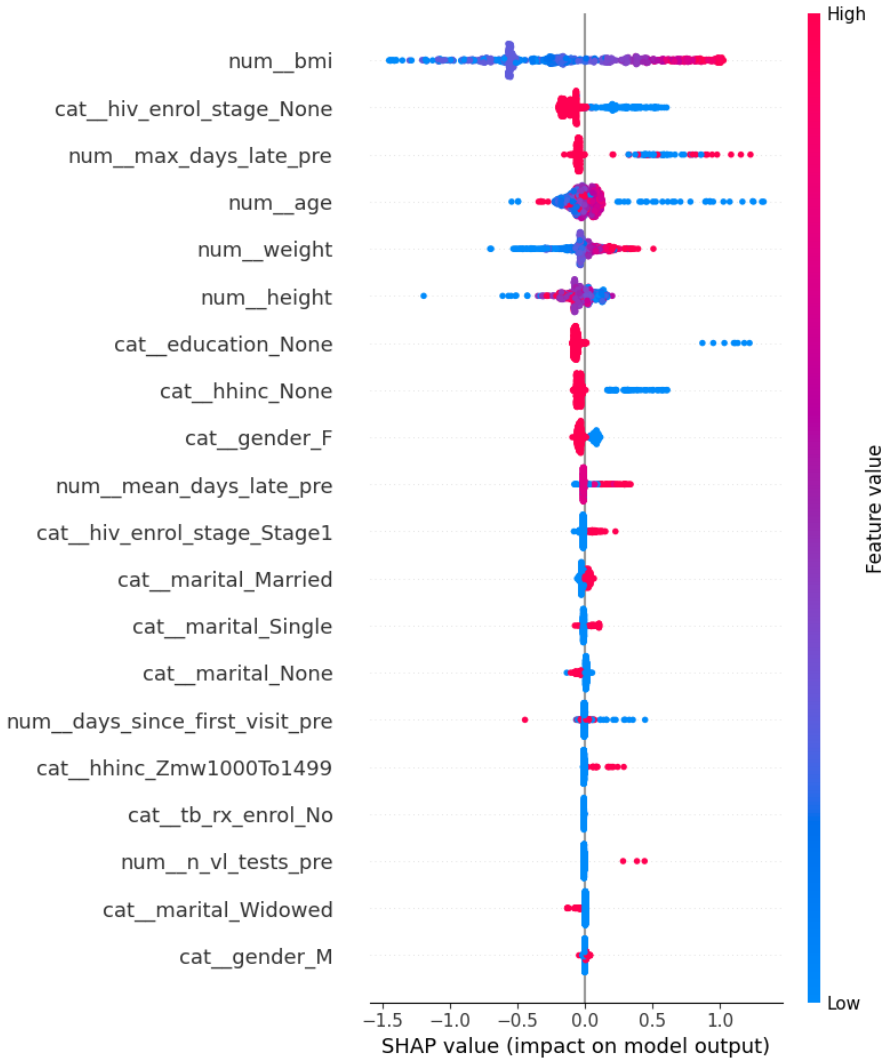

Figure 5.13: *SHAP global summary (beeswarm) for TB12m. The BMI cloud separates cleanly by colour — low-BMI patients (blue) on the risk-increasing side, high-BMI patients (red) on the protective side — confirming that the model uses BMI in the clinically correct direction.*

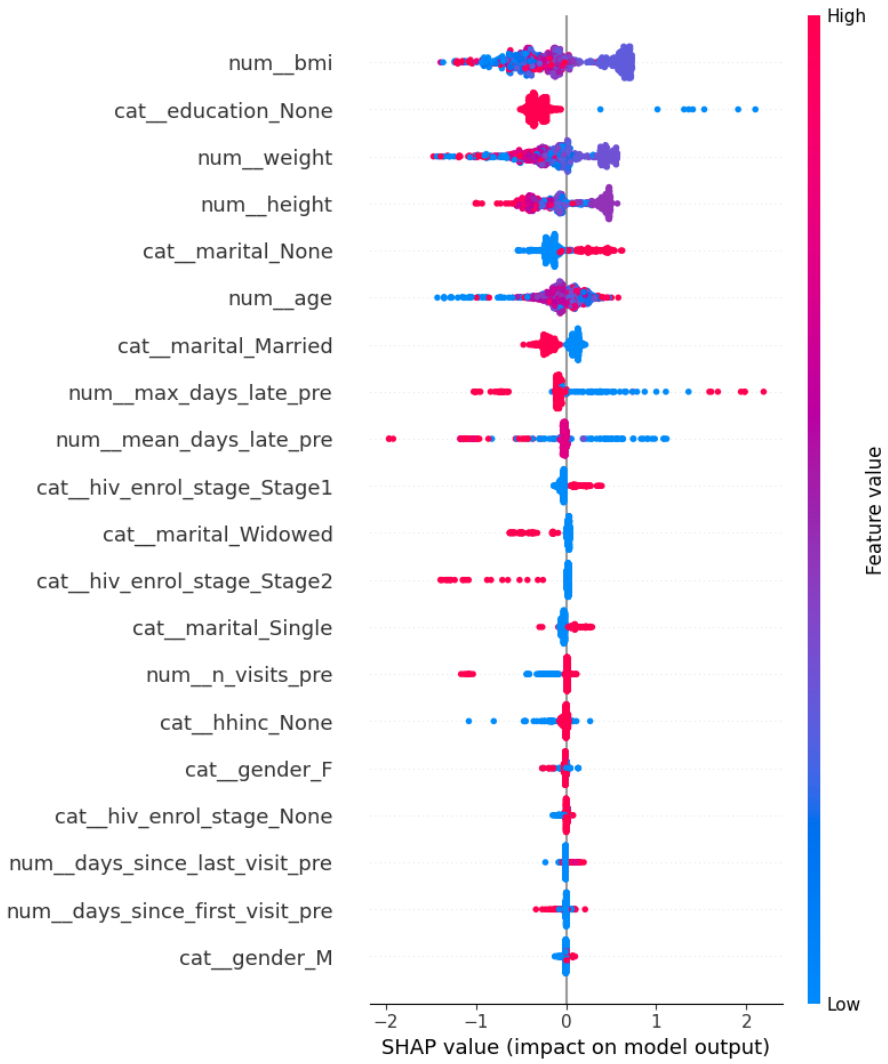

Figure 5.14: *SHAP global summary (beeswarm) for IIT12m. Engagement features lead; high lateness values (red) push predictions toward higher interruption risk. The compactness of most clouds around zero reflects the extreme class imbalance and a model that confidently isolates a small high-risk group.*

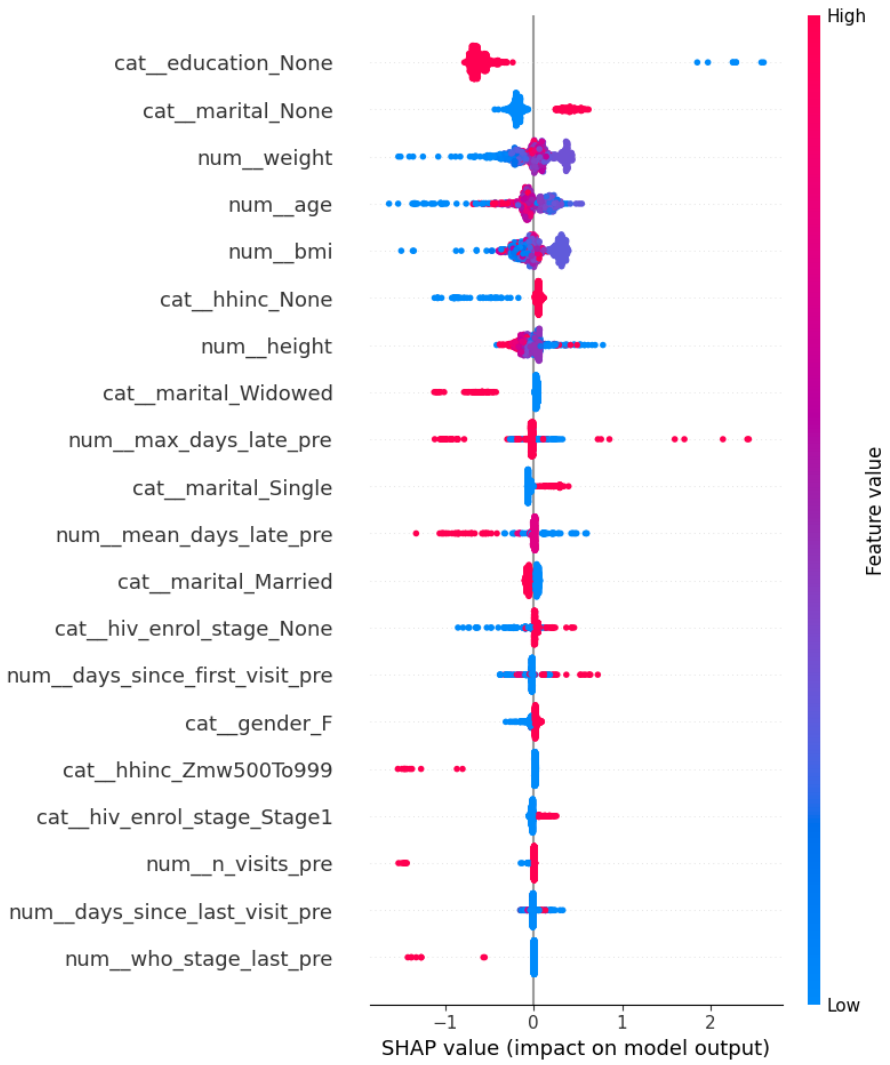

Figure 5.15: *SHAP global summary (beeswarm) for UVL12m. The encoded absence of an education record (cat__education_None) is the top feature and its value-dependence is the inverse of a clinical gradient — the visual signature of the missingness-driven prediction pattern that the thesis treats as a deployment risk requiring human-in-the-loop review.*

Figure 5.15 is, for the equity argument of this thesis, the single most consequential result of the explainability analysis. The dominance of an encoded missingness indicator at the top of the UVL12m beeswarm demonstrates that for this outcome the model predicts, to a meaningful degree, on the basis of whether an education record exists rather than what it contains. Because education completeness is lowest at the facilities serving the most deprived populations, an unguarded deployment acting on these scores could systematically mis-prioritise patients along lines correlated with deprivation. This empirical observation is the foundation for the mandatory-review

provision and the fairness-audit approval gate specified in the deployment and governance architectures of Chapters 6 and 7.

5.4.3 SHAP Local (Patient-Level) Explanations

SHAP waterfall plots were generated for three representative patients per outcome (highest-risk, median-risk and lowest-risk) on the held-out test set. The complete graphical waterfall plots produced directly by the analytical pipeline are reproduced in full as Figures 20 to 28 below, organised by outcome (TB12m, then IIT12m, then UVL12m) and within each outcome by ascending risk (lowest, median, highest). Each plot decomposes a single patient's predicted log-odds into additive feature contributions, beginning from the model base value E[f(X)] and accumulating to the patient-specific prediction f(x); red bars indicate features that increase predicted risk and blue bars indicate features that decrease it. The corresponding summary tables of top contributions are provided in Appendix C. Key findings are summarised below.

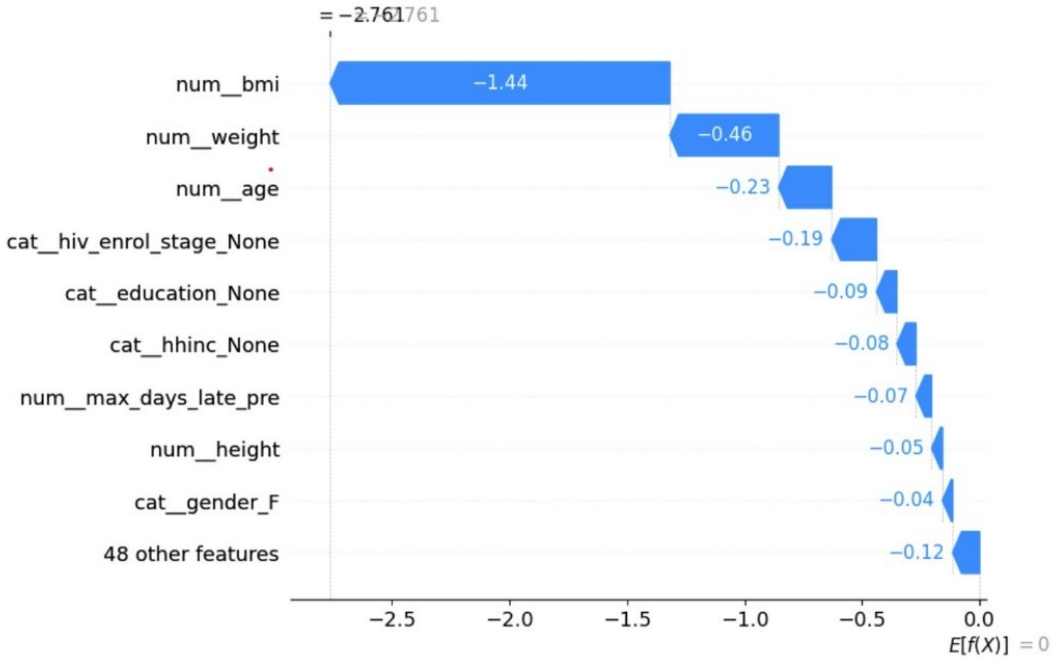

Figure 5.16: *TB12m, lowest-risk representative patient (predicted P ≈ 0.059). All contributions are negative (blue), driving the prediction far below the model base value. BMI is the dominant protective feature (−1.44), followed by weight and age, reflecting a nutritionally adequate younger profile with complete clinical documentation. Final log-odds f(x) = −2.761.*

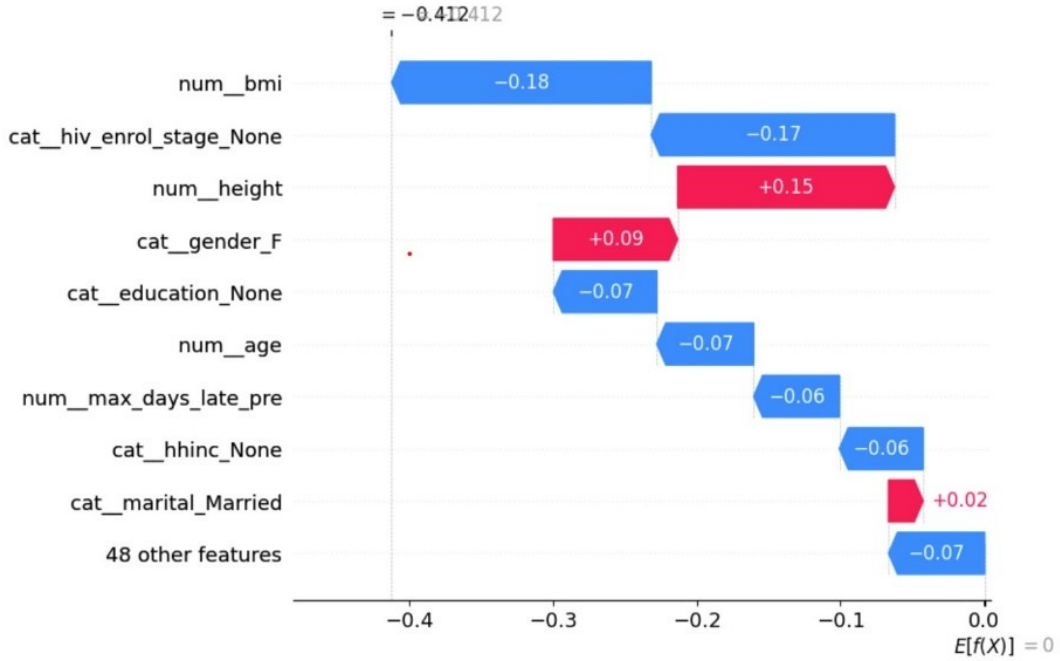

Figure 5.17: *TB12m, median-risk representative patient (predicted P ≈ 0.398). Competing protective (blue) and risk-increasing (red) contributions produce a near-baseline prediction. BMI and missing HIV enrolment stage are protective; height and female sex are marginally risk-increasing. This is the typical clinical decision context in which a risk score alone is insufficient. Final log-odds f(x) = −0.412.*

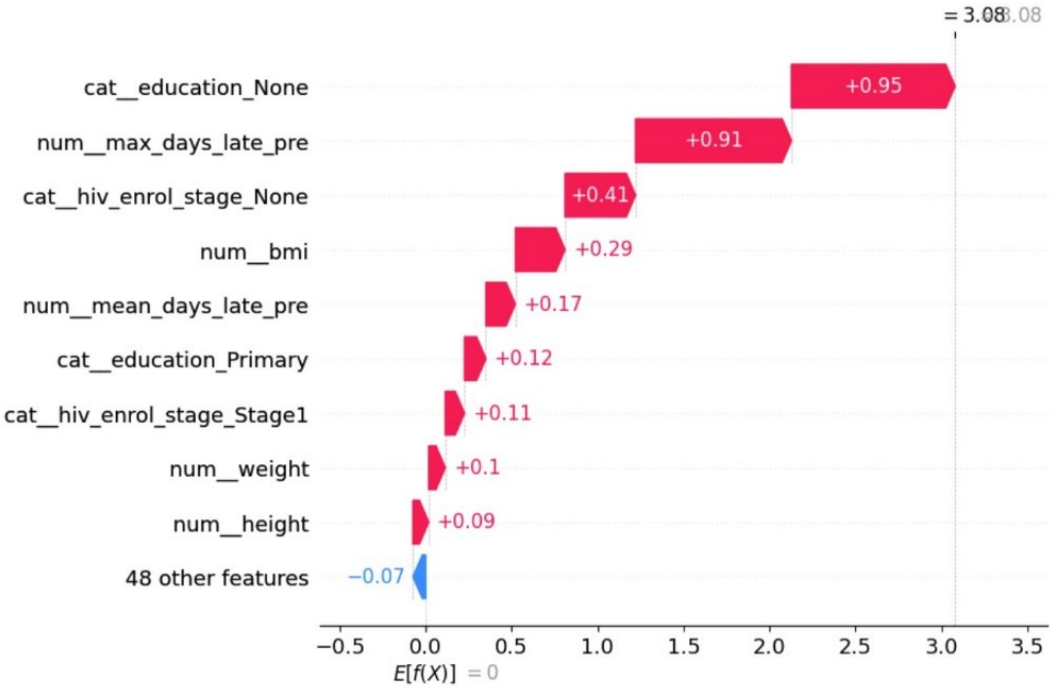

Figure 5.18: *TB12m, highest-risk representative patient (predicted P ≈ 0.956). The two largest contributions are absence of recorded education (cat__education_None, +0.95) and extreme appointment lateness*

(num__max_days_late_pre, +0.91), ahead of missing HIV enrolment stage (+0.41). For this individual, engagement and documentation gaps, not anthropometric features, are the primary risk drivers, pointing to outreach and documentation follow-up rather than nutritional intervention. Final log-odds f(x) = +3.08.

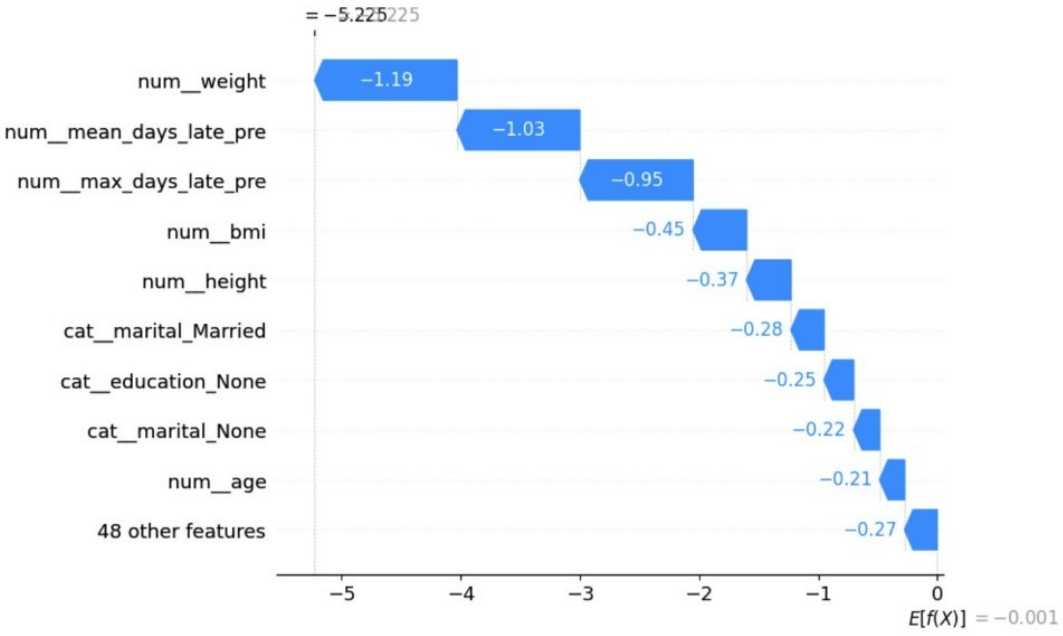

Figure 5.19: *IIT12m, lowest-risk representative patient (predicted P ≈ 0.005). A virologically suppressed, well-engaged profile yields strongly negative contributions across viral-load and visit-count features. Final log-odds well below baseline.*

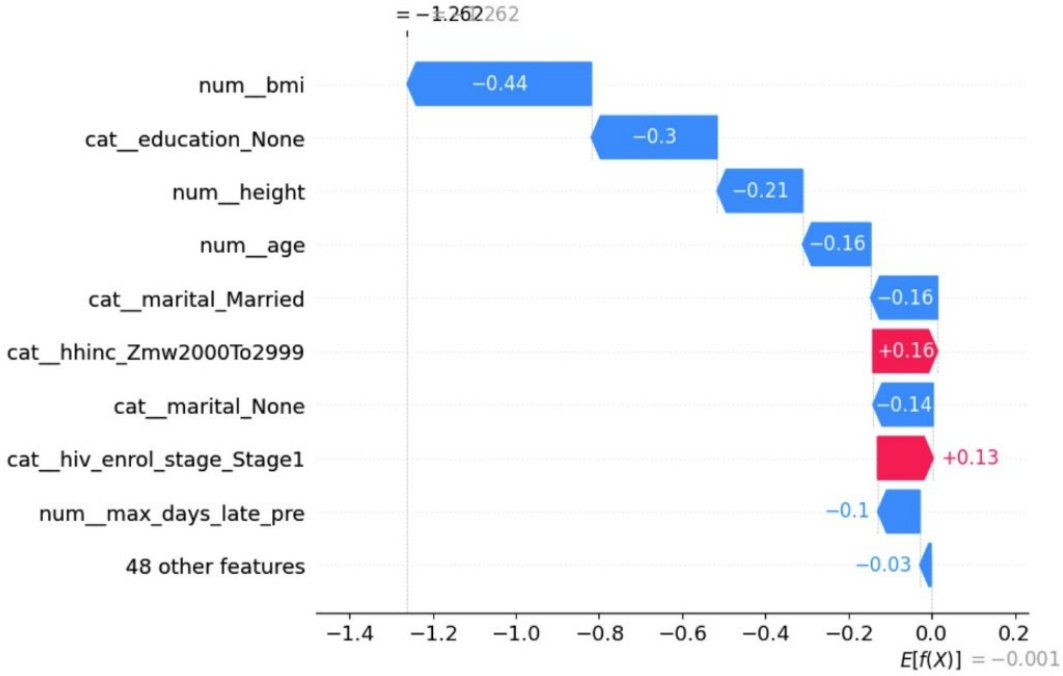

Figure 5.20: *IIT12m, median-risk representative patient (predicted P ≈ 0.221). Mixed contributions dominated by engagement and documentation features; missing education and marital status provide modest positive contributions while anthropometric features are weakly protective.*

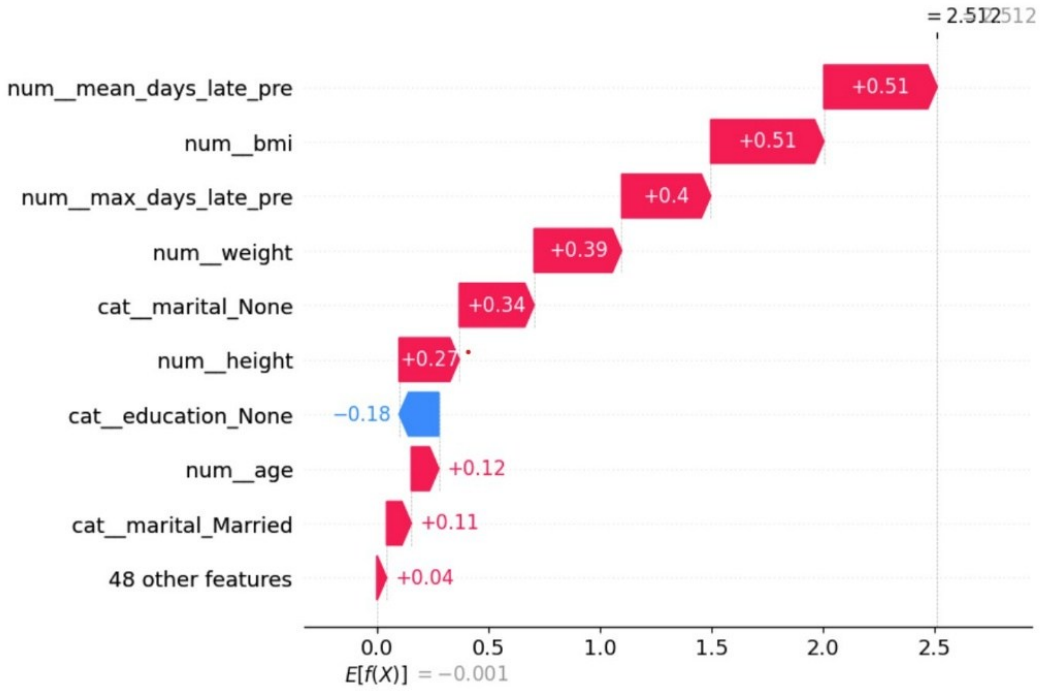

Figure 5.21: *IIT12m, highest-risk representative patient (predicted P ≈ 0.925). Dominated by engagement signals (mean and maximum days late) and anthropometric signal (BMI), all clinically interpretable and actionable. In contrast to the TB12m highest-risk case, this rationale is engagement-driven rather than documentation-driven, illustrating how the same framework produces different rationales for different patients.*

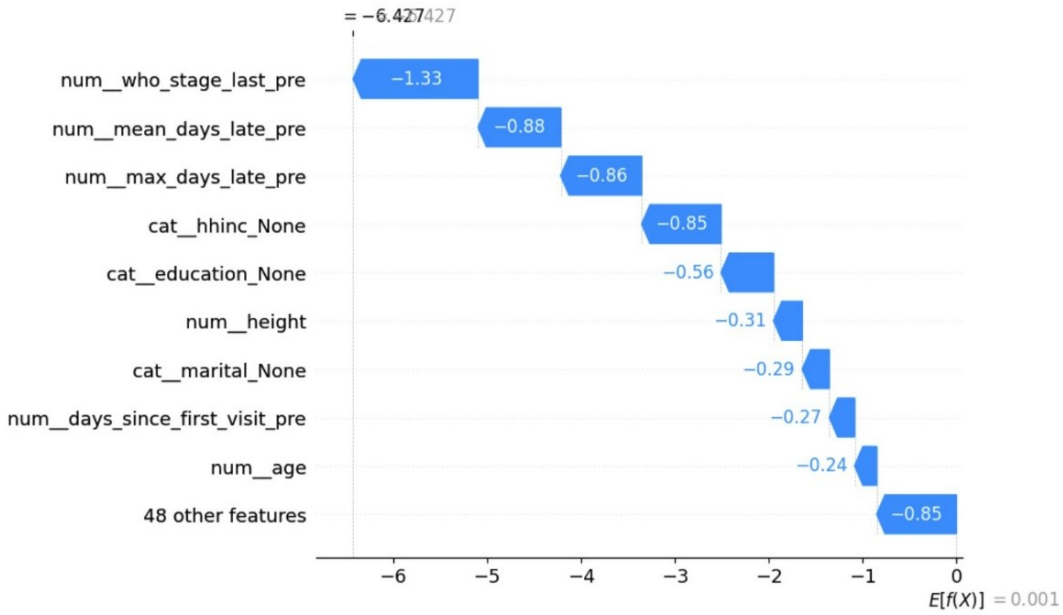

Figure 5.22: *UVL12m, lowest-risk representative patient (predicted P ≈ 0.002). Recorded tertiary education and viral suppression contribute strongly protective values; the model has learnt an education gradient consistent with the health-literacy literature. Final log-odds far below baseline.*

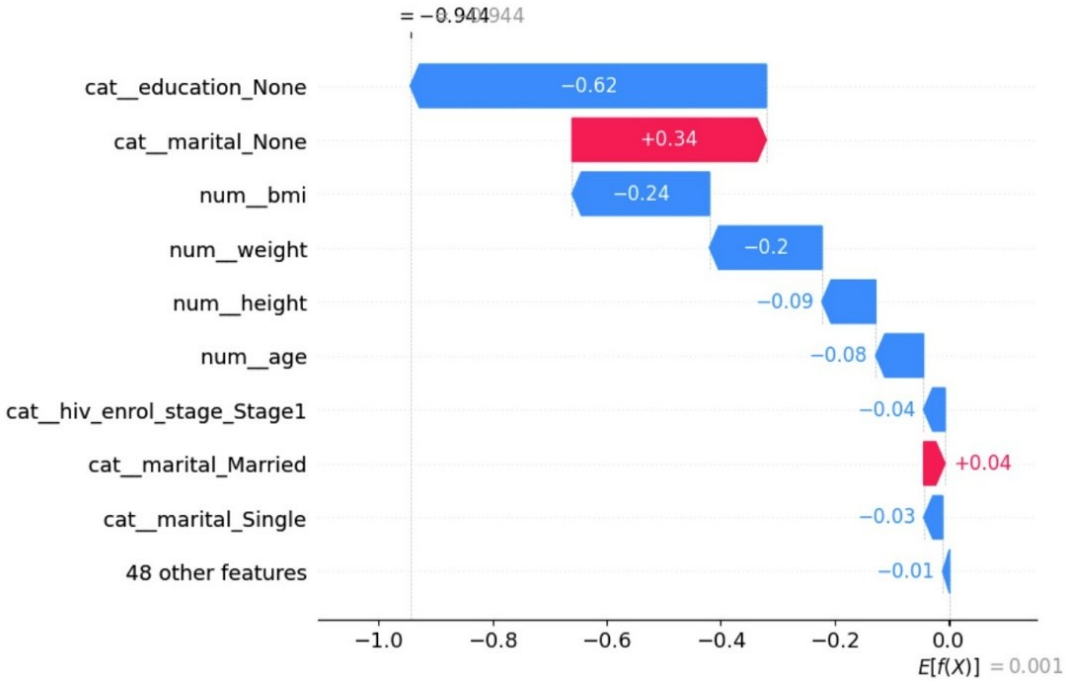

Figure 5.23: *UVL12m, median-risk representative patient (predicted P ≈ 0.280). Missing income and demographic features provide small positive contributions; anthropometric features are weakly protective, producing a moderate prediction.*

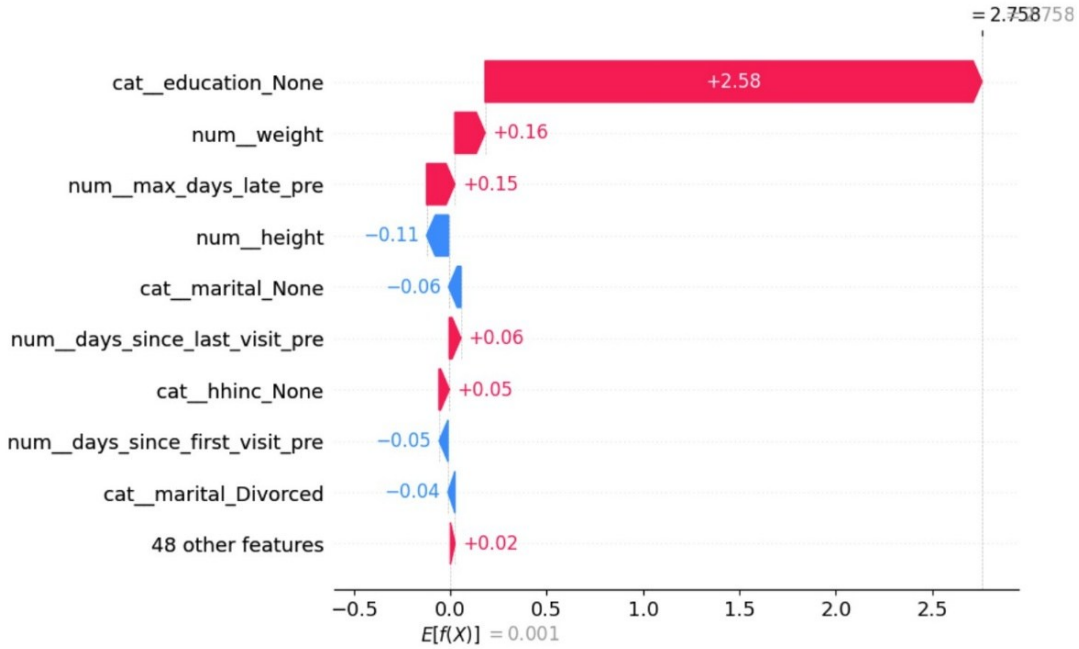

Figure 5.24: *UVL12m, highest-risk representative patient (predicted P ≈ 0.940). The most extreme single-feature dominance observed in the study: cat__education_None alone contributes +2.58 to the log-odds, with all other contributions minor. Combined with the missingness-outcome confounding finding (Section 5.5.7), this provides the strongest empirical case for the human-in-the-loop review provision in the proposed deployment architecture. Final log-odds f(x) = +2.758.*

TB12m highest-risk patient (f(x) = +3.08; P ≈ 0.956). The two largest contributions were absence of recorded education (+0.95) and extreme appointment lateness (+0.91), ahead of missing HIV enrolment stage (+0.41). While BMI dominates at the population level, engagement and documentation gaps can be the primary individual-level risk signal.

IIT12m highest-risk patient (f(x) = +2.512; P ≈ 0.925). Driven by mean days late (+0.51), BMI (+0.51), maximum days late (+0.40) and weight (+0.39). The dominant features are engagement and anthropometric, not demographic.

UVL12m highest-risk patient (f(x) = +2.758; P ≈ 0.940). Almost entirely explained by a single feature: education_None (+2.58). This represents the most extreme single-feature dominance observed in the study and makes the clinical recommendation unambiguous: the patient's risk is primarily structural and addressable through health literacy and adherence support — but interpretation must be guarded by the missingness-outcome confounding finding (Section 5.5.7).

These local explanations show that the framework provides not a risk score alone but a patient-specific, decomposed rationale, a prerequisite for clinician trust and targeted intervention in routine HIV care. The complete representative-patient decomposition tables for all nine patients are presented in Appendix C.

5.4.4 Cross-Facility Generalisability

Consistent performance on the two held-out Matero facilities, which contributed no data to training, provides evidence of cross-facility generalisability. This is particularly notable for TB12m, where test-set prevalence (41.3%) substantially exceeded training-set prevalence (33.7%), yet AUC-PR remained high (0.625). For IIT12m and UVL12m, the low PR-AUC values on both training and test subsets reflect inherent signal limitation in the available pre-index EHR features, rather than generalisation failure.

5.5 Study Three: Data Quality Assessment Findings

5.5.1 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). Enrolments spanned 2003 to 2025. Patient volumes per facility ranged from 26,845 (Matero Main UHC) to 56,871 (Kanyama FLH), a 2.1-fold difference. All six facilities scored below the 70% DQI threshold, confirming that the data quality challenges identified are programme-wide rather than facility-specific. The preprocessing code required 23 separate conditional checks to handle cases where data fields were present in some facilities but absent or differently named in others; each check is a signal of cross-facility schema inconsistency.

The cross-facility heterogeneity has direct implications for AI deployment. A model trained on one facility's schema cannot be deployed at another without facility-specific data mapping, and the data quality differences across facilities mean that the same model may produce systematically different predictions when applied to different facilities. The proposed CDSS integration architecture (Chapter 6) addresses this by enforcing schema standardisation through the FHIR Adapter Service, but the empirical observation here is that the underlying schema heterogeneity is itself a programme-level governance issue that requires resolution.

5.5.2 Field Completeness: Three-Tier Pattern (D1)

Field completeness rates across 15 AI-critical features are presented in Table 5.13 and illustrated in Figure 9. Three distinct tiers emerge from the data.

High completeness (≥80%): Basic registration fields (gender, age, TB treatment at enrolment: 99–100%) and physical measurements (height: 72–90%; weight: 84–93%; BMI: 70–88%). These fields are reliably recorded across all six facilities.

Moderate completeness (40–79%): Marital status (70–80%). HIV enrolment stage is recorded for only 21–25% of patients, a concerning finding because it is the third most important predictor of TB incidence in the companion ML model.

Critical missingness (<10%): Education level (1.2–7.9%), household income (4.7–19.8%), past TB type (0.5–1.3%), exit reason and last interruption date (0% at every facility). The 0%

completeness of exit reason and last IIT date is particularly striking, as these fields are operationally important for the IIT12m prediction.

Table 5.13: *Per-field, per-facility EHR completeness rates (%) for 15 key AI features.*

Field	Chawama	Chelstone	George	Kanyama	Matero Ref	Matero Main	Mean
Gender	99	100	100	100	100	100	100
Age	99	100	100	100	100	100	100
TB Rx Enrolment	100	100	100	100	100	100	100
Weight	89	93	88	93	88	84	89
Height	85	89	86	89	86	72	85
BMI	84	88	84	88	84	70	83
Marital Status	78	74	70	80	70	80	75
HIV Enrol. Stage	24	22	25	23	21	21	23
HIV Enrol. 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 IIT Date	0	0	0	0	0	0	0

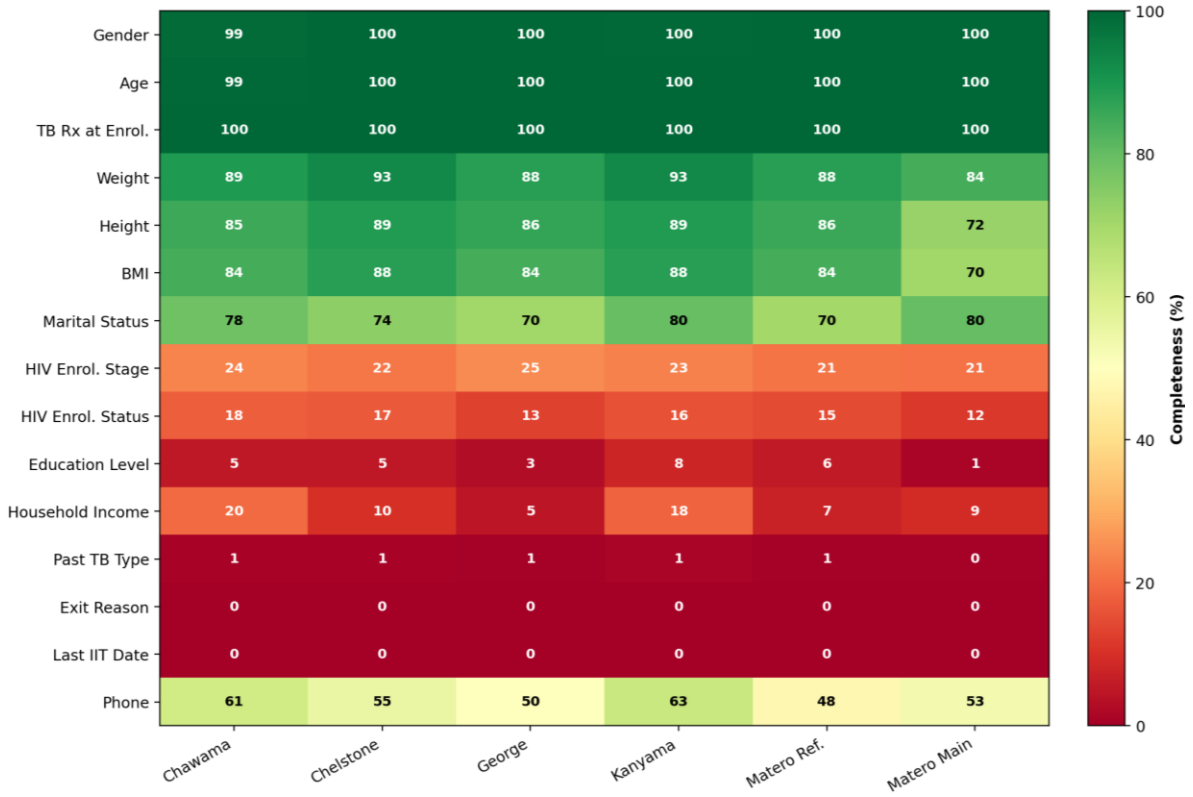

Figure 5.25: *EHR field completeness heatmap across six Lusaka HIV facilities (red = 0%; green = 100%). Education level and household income appear as uniformly red columns, effectively absent from the digital record system. The three-tier completeness pattern is clearly visible.*

5.5.3 Feature Coverage Collapse (D4)

The most consequential data quality finding concerns feature coverage, the proportion of patients for whom each feature group can be computed from pre-index data. As shown in Figure 5.26 and Table 5.14:

- **Demographic features (File 1):** 100% coverage, comprising all 246,053 patients.
- **Engagement/visit features (File 2):** Only 12.3% coverage, comprising 30,295 of 246,053 patients (counting distinct facility–patient records, the unit on which feature construction keys; 222 client identifiers occur at more than one facility). The remaining 87.7% received zero-imputed engagement features.
- **Laboratory features (File 4):** Only 3.3% coverage, comprising 8,233 of 246,053 patients. The remaining 96.7% received zero-imputed laboratory features.

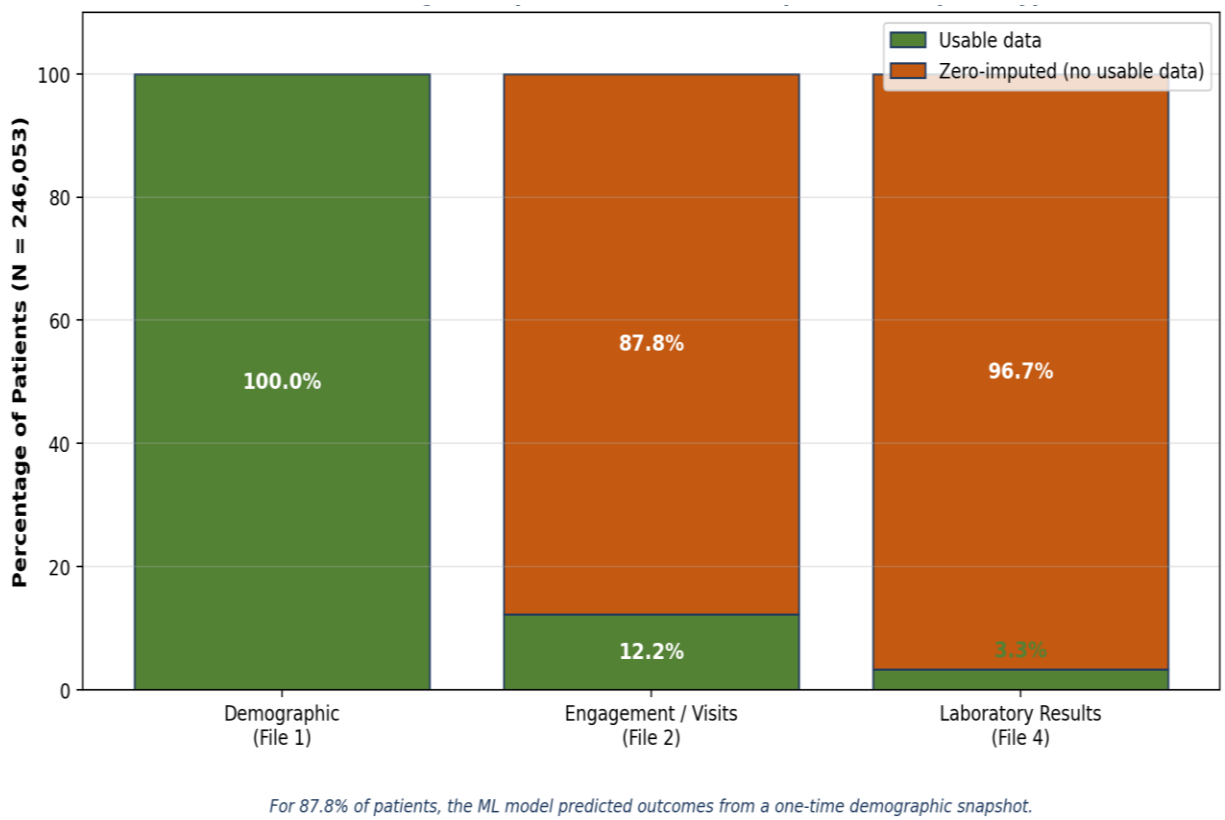

Figure 5.26: *Feature coverage collapse. Only 12.3% of patients had usable engagement data and 3.3% had usable laboratory data; the remainder were zero-imputed. For 87.8% of patients, the ML model predicted outcomes from a one-time demographic snapshot.*

Table 5.14: *Feature coverage by data file type.*

Feature Type	Total Patients	Patients with Usable Data	Coverage %	Zero-imputed %
Demographic (File 1)	246,053	246,053	100.0%	0.0%
Engagement/Visits (File 2)	246,053	30,295	12.3%	87.7%
Laboratory Results (File 4)	246,053	8,233	3.3%	96.7%

This finding is, to the author's knowledge, the first systematic quantification of feature coverage collapse in an LMIC HIV EHR context. The standard narrative in the published ML-in-LMIC literature is that models are trained on multi-modal longitudinal data; the reality, exposed by this analysis, is that for 87.8% of patients in this Zambian cohort the model effectively predicted from

a one-time demographic and anthropometric snapshot at enrolment. The implications for model interpretation are discussed extensively in Chapter 6 (Section 6.3).

5.5.4 Plausibility Violations (D3)

Among the 30,295 patients with engagement data, 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, indicating data entry errors in the appointment scheduling system. These extreme values entered the ML model without any plausibility check, and appointment lateness is among the top-five ML features for both TB and treatment interruption prediction. Table 5.15 (referenced below in Section 5.5.5) reports the plausibility violation counts.

Table 5.15: *Plausibility violation counts and rates among patients with engagement data (n = 30,295).*

Violation Type	N Violations	% of Eligible Patients
Max appointment lateness >730 days	16,439	54.7%
Max appointment lateness >365 days	12,103	40.3%
Negative lateness >30 days	3,965	13.2%
Maximum recorded lateness (single patient)	45,459 days	~124 years
BMI >60 kg/m²	847	2.8%
BMI <10 kg/m²	152	0.5%

5.5.5 Outcome Sensitivity (D5)

Outcome prevalence varied substantially across facilities and across train-test subsets, as shown in Table 20. The TB12m facility range (32.1–48.4%) reflects genuine epidemiological heterogeneity; the IIT12m range (0.20–0.65%) and the UVL12m range (0.80–2.19%) include both real variation and operational variation in coding and testing intensity.

Table 5.16: *Outcome prevalence across data subsets and facilities.*

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%
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%	Range: 0.20%-0.65%; inconsistent coding
IIT12m	Training set	905	0.53%	—
IIT12m	Test set	473	0.63%	—
UVL12m	All facilities	3,564	1.45%	Range: 0.80%-2.19%; differential testing
UVL12m	Training set	2,693	1.57%	—
UVL12m	Test set	871	1.17%	—

A specific data quality concern arises for TB12m: the 36% overall prevalence substantially exceeds typical TB incidence in PLHIV cohorts. This reflects label conflation: the data fields available do not reliably distinguish prevalent TB (present at index) from incident TB (newly diagnosed within 12 months), so the operationalised label includes both. This is itself a substantive DQA finding, requiring schema improvements at the SmartCare data entry level to support cleaner outcome labelling.

5.5.6 Temporal Stability (D6)

Field completeness stratified by patient enrolment year revealed three key patterns. First, basic registration data has been recorded reliably throughout the programme's history (1999–2025). Second, HIV enrolment stage and household income completeness briefly improved in 2016–2019 before declining, suggesting a data entry campaign that was not sustained. Third, education completeness has never exceeded approximately 8% in any enrolment year since 2004; the persistent missing-data pattern for education represents an ongoing structural gap in data collection

practice. These findings have direct implications for any future data quality remediation: short-term campaigns alone are insufficient; sustained operational change is needed.

5.5.7 Missing-Data Confounding Test (D8)

Pearson correlations between facility-level missingness rates and outcome rates were computed for nine high-importance feature–outcome pairs. This analysis is exploratory, ecological and hypothesis-generating: it is computed across only six facility-level data points, has limited statistical power, cannot detect within-facility effects, and is unadjusted for case mix. The $|r| > 0.5$ line is therefore used as a screening flag for review rather than a confirmatory threshold, and no causal claim is made from it. Results are presented in Table 5.17; the patient-level analysis that follows (Table 5.18) is the stronger evidence.

Table 5.17: *Missing-data confounding test results. $|r| > 0.5$ used as a screening flag (n = 6 facilities; exploratory, ecological).*

Feature (missing-data indicator)	Outcome	Correlation r	Interpretation
Education	TB12m	+0.599	Flag for review — facilities with more missing education records also record higher TB-treatment rates
Education	IIT12m	+0.105	Low risk
Education	UVL12m	−0.865	Flag for review — facilities with less complete education data record fewer unsuppressed-VL events
Household income	TB12m	+0.263	Low risk
Household income	UVL12m	−0.478	Borderline — investigate
HIV enrolment stage	TB12m	+0.741	Flag for review — facilities with more missing stage records record higher TB-treatment rates
HIV enrolment stage	IIT12m	−0.085	Low risk
HIV enrolment stage	UVL12m	+0.126	Low risk

Feature (missing-data indicator)	Outcome	Correlation r	Interpretation
Marital status	TB12m	−0.429	Low risk

Three feature–outcome pairs exceed the |r| > 0.5 screening line: education–TB (+0.599), education–UVL (−0.865) and stage–TB (+0.741). Because these are ecological correlations over six facilities, they are treated as flags for the patient-level investigation reported next rather than as established confounding, and their implications for model trust are discussed, with the necessary confirmatory analyses specified, in Section 6.3.

The facility-level analysis above is necessarily coarse. To interrogate the missingness at the level of the individual patient, the cohort was partitioned by whether the education field — the most important model feature and the most severely incomplete — was recorded (n = 12,766) or missing (n = 233,287), and the two groups compared on the better-recorded fields and on outcome rates. The contrast is stark and is the strongest single piece of evidence that missingness in this dataset is not random. Relative to the missing-education group, the recorded-education group has roughly three times the pre-index engagement-data availability (34.7% versus 11.1%), about twenty times the pre-index CD4 or viral-load value availability (12.5% versus 0.6%), and much higher recorded-outcome rates (recorded TB treatment 59.0% versus 34.7%; programme-recorded interruption 3.24% versus 0.41%; recorded unsuppressed viral load 12.11% versus 0.87%), despite near-identical median age (31 versus 32 years) and sex distribution (63.3% versus 62.4% female). Table 5.18 reports the comparison.

Table 5.18: Patient-level comparison of patients whose education field is recorded versus missing, on better-recorded fields and outcome rates.

Characteristic	Education recorded (n = 12,766)	Education missing (n = 233,287)
Median age (years)	31	32
Female	63.3%	62.4%
Pre-index engagement-data availability	34.7%	11.1%
Pre-index CD4 / viral-load value availability	12.5%	0.6%
Recorded TB treatment (TB12m)	59.0%	34.7%
Programme-recorded interruption (IIT12m)	3.24%	0.41%
Recorded unsuppressed viral load (UVL12m)	12.11%	0.87%

The recorded-education subgroup is therefore demonstrably not representative of the cohort: it is more completely documented on every axis and has far higher recorded-event rates, while being

demographically almost identical. This co-variation of missingness with the recorded outcomes is precisely what a facility-level correlation cannot show and an imputation would conceal, and it is why the framework treats an explicit observed-versus-missing comparison, rather than imputation, as the appropriate response to non-random missingness. Because the underlying joint patient-level distributions cannot be released under the data-governance terms, this comparison is regenerated by the released code on the restricted dataset.

5.5.8 Composite Data Quality Index

The composite DQI per facility ranged from 53.5% (Matero Main Urban Health Centre) to 58.3% (Kanyama First Level Hospital), as shown in Figure 5.27 and Table 23. No facility exceeded the 70% minimum threshold. The narrow gap between best and worst facility (4.8 percentage points) confirms a programme-wide challenge requiring programme-level governance response rather than facility-specific intervention.

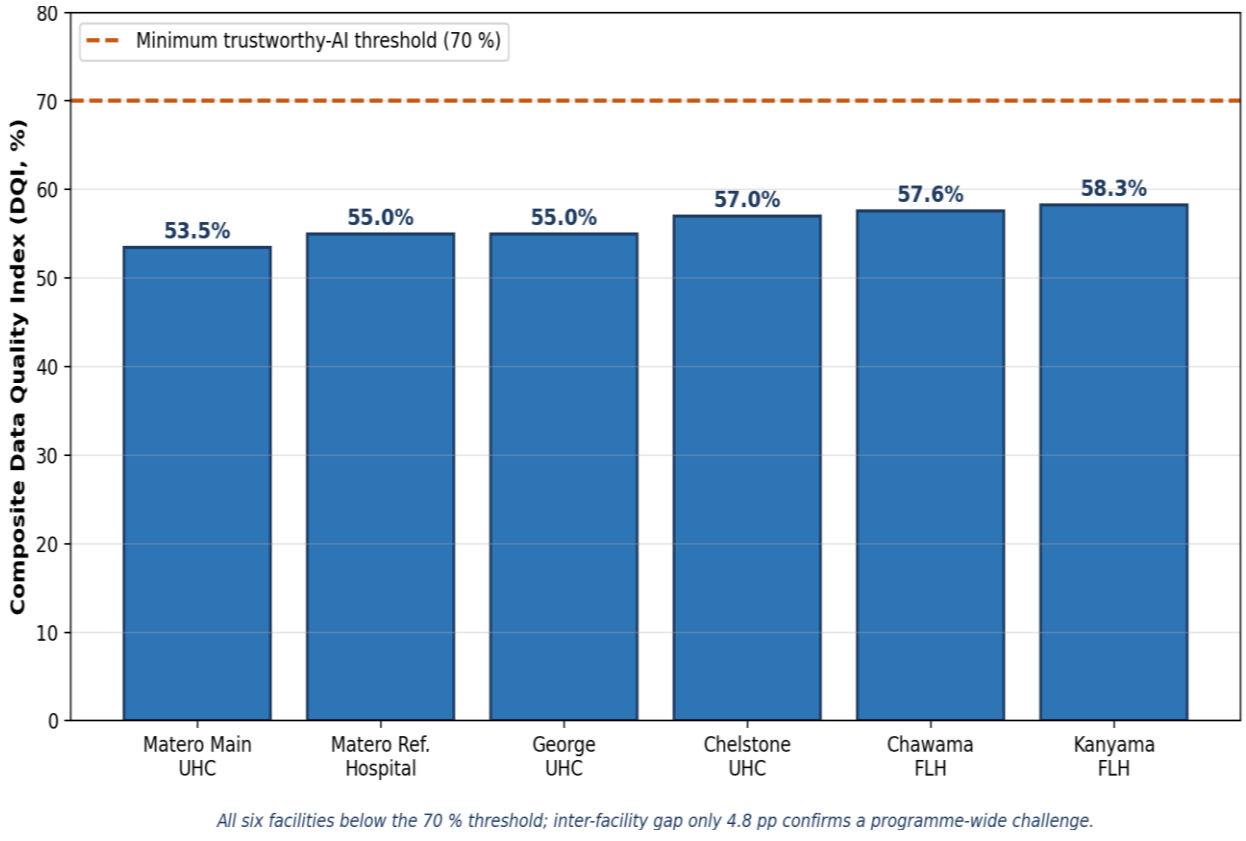

Figure 5.27: *Composite Data Quality Index (DQI) by facility. All six facilities score below the 70% minimum threshold for reliable AI training, confirming a programme-wide data quality challenge. The gap between the best (58.3%) and worst (53.5%) facility is only 4.8 percentage points.*

Table 5.19: *Facility-level composite DQI decomposition (average completeness across 15 AI-critical fields).*

Facility	DQI (%)	Best-Performing Field	Weakest-Performing Field
Kanyama First Level Hospital	58.3	Gender / Age / TB Rx (100%)	Past TB Type (1.3%)
Chawama First Level Hospital	57.6	Gender / Age / TB Rx (~99%)	Education (5.1%)
Chelstone Urban Health Centre	57.0	TB Rx (100%)	Education (5.4%)
George Urban Health Centre	55.0	TB Rx (100%)	Education (3.0%)
Matero Reference Hospital	55.0	TB Rx (100%)	Past TB Type (0.6%)
Matero Main Urban Health Centre	53.5	TB Rx (100%)	Education (1.2%)

5.5.9 Composite Index Against Trustworthiness Thresholds

Figure 5.28 plots the composite Data Quality Index for each facility against the two governance thresholds proposed in this thesis: a 50% minimum-usability line below which data are considered unfit for any automated decision support, and a 70% trustworthy-AI line above which fully automated (non-shadow) deployment could be contemplated. No facility reaches the 70% line, and all sit between the two thresholds, which is the single most important data-governance finding of the study.

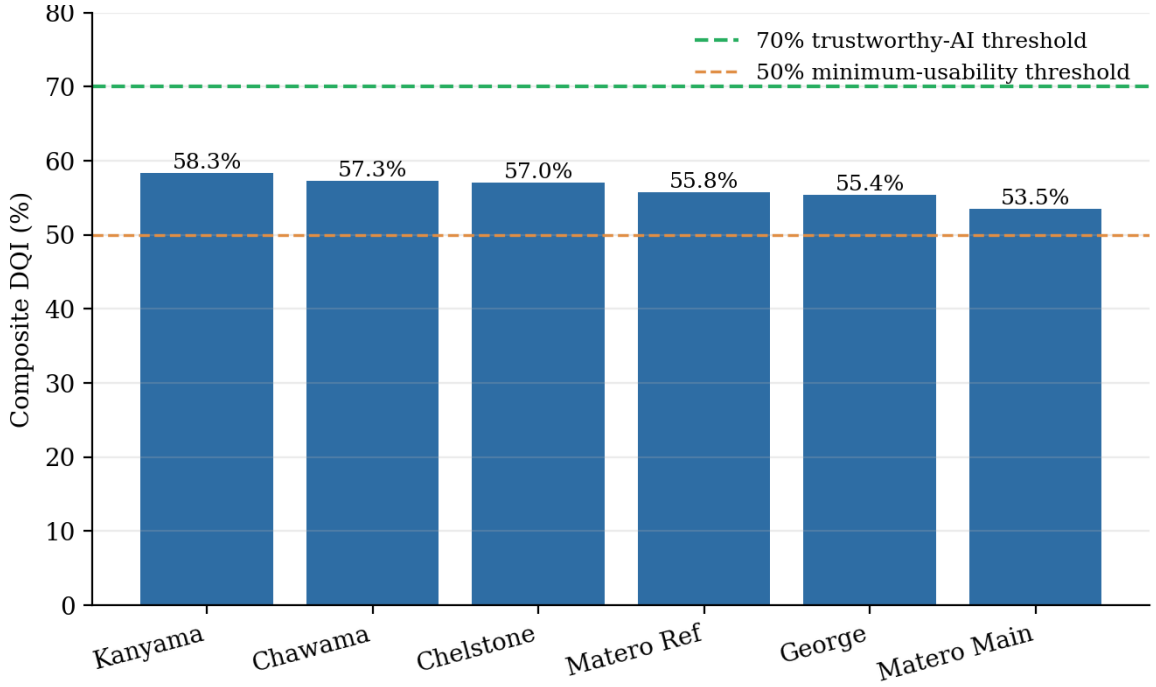

Figure 5.28: *Composite Data Quality Index by facility against the 50% minimum-usability and 70% trustworthy-AI thresholds. All six facilities fall in the intermediate band (53.5%–58.3%), supporting a shadow-mode deployment with mandatory human review rather than full automation.*

Table 5.20: *Summary of the eight-dimension data-quality assessment framework applied to the 15 ML-critical fields, with the headline finding and modelling implication for each dimension.*

Dimension	What it measures	Headline finding	Implication
D1 Completeness	Proportion of non-missing values per field	Three-tier structure; predictive fields 1–20% complete	Drives missingness-imputation strategy
D2 Conformity	Values within expected format/range	Implausible lateness up to 45,459 days	Requires plausibility edit-checks
D3 Plausibility	Clinical/biological believability	BMI and date outliers present	Winsorisation before modelling
D4 Consistency	Agreement across related fields	Visit and outcome dates internally consistent	Low risk to outcome labels
D5 Uniqueness	Absence of duplicate records	Client IDs unique within facility	Supports patient-level analysis
D6 Temporal stability	Stability of completeness over time	Episodic, campaign-driven collection	Cohort-era confounding risk

Dimension	What it measures	Headline finding	Implication
D7 Timeliness	Currency of records relative to index	Exit reason and last-IIT-date empty	Limits some outcome definitions
D8 Representativeness	Coverage across facilities/groups	Uniform DQI across six facilities	Programme-wide, not facility-specific, remediation

Table 5.20 consolidates the eight-dimension assessment that, to the best of available evidence, has not previously been applied in full to a Zambian HIV EHR for the explicit purpose of certifying machine-learning readiness. The framework's contribution is to move the data-quality discussion beyond simple completeness percentages toward an auditable, multi-dimensional certification that maps each deficit to a concrete modelling or governance action. The dominant pattern — acceptable consistency and uniqueness but poor completeness, conformity and temporal stability in the most predictive fields — is precisely the profile that motivates the shadow-mode, human-in-the-loop deployment recommended in Chapter 6 rather than immediate full automation.

5.6 Cross-Study Synthesis

This section synthesises findings across the empirical studies and links them to the four-component framework introduced in Chapter 2. Table 5.12 presents the cross-study synthesis matrix.

Table 5.21: *Cross-study synthesis matrix: findings, contributions and links to framework components.*

Finding	Source Study	Component Addressed	Implication
Only 18.8% LMIC studies; 7.8% external validation; 6.2% calibration	Study One	Methodological	Justifies facility-holdout design + calibration as standard
Education completeness 1.2–7.9% across facilities	Study Three	Data Quality	Programme-wide intervention required, not facility-specific
Feature coverage 12.3% engagement, 3.3% lab	Study Three	Data Quality	87.8% predicted from demographic snapshot alone

Finding	Source Study	Component Addressed	Implication
54.3% implausible appointment lateness	Study Three	Data Quality	Plausibility gating mandatory in CDSS
Missingness-outcome r = −0.865 (education-UVL)	Study Three	Equity / Governance	Equity safeguards required in deployment
AUC-ROC TB12m 0.707; IIT12m 0.839; UVL12m 0.778	Study Two	Methodological	Clinically meaningful discrimination achievable
Education dominates UVL12m SHAP & permutation	Study Two + Three	Methodological + Equity	Confounded signal: human-in-loop review required
F1-optimal τ* enables outcome-specific protocols	Study Two	Methodological	Differentiated operational deployment by outcome
Facility-holdout validation succeeds despite shift	Study Two	Methodological	Cross-site deployment plausible
Stakeholder endorsement at three dissemination meetings	Dissemination	Governance	Governance pathway is feasible
No facility ≥ 70% DQI; range 53.5–58.3%	Study Three	Data Quality + Governance	DQI ≥ 70% must be gate criterion for deployment

Three cross-cutting patterns emerge from this synthesis. First, methodological gaps identified in Study One are addressed in Study Two through deliberate pipeline design (facility-holdout, leakage-safe, calibration, SHAP), showing that the gaps are tractable when prioritised. Second, the data quality findings of Study Three do not invalidate the ML findings of Study Two, but they qualify them in operationally consequential ways, particularly through the missingness-outcome confounding result that interprets the dominance of education as a partially-confounded signal. Third, the empirical findings of all studies converge on the same governance and deployment requirements: data quality gating, equity safeguards, human-in-the-loop review and explainability as a precondition for deployment.

5.7 Subgroup Performance Analysis

The empirical performance of the ML framework was further analysed across patient subgroups defined by sex, age band, education level (where recorded) and facility. Subgroup AUC-ROC and AUC-PR are reported for the stacked ensemble model in this section, providing the empirical foundation for the fairness safeguards described in Chapter 6. A subgroup performance gap of more than 0.10 in AUC-ROC between any two subgroups is flagged as a fairness concern requiring governance review.

5.7.1 Subgroup Performance by Sex

Performance was analysed separately for female patients (n=46,566 in the test set) and male patients (n=27,940 in the test set). Table 5.21 presents the subgroup AUC-ROC and AUC-PR for the stacked ensemble model across the three outcomes.

Table 5.22: *Stacked ensemble performance on the held-out test set by sex.*

Outcome	Female AUC-ROC	Male AUC-ROC	Gap	Female AUC-PR	Male AUC-PR	Gap
TB12m	0.704	0.711	0.007	0.617	0.638	0.021
IIT12m	0.832	0.840	0.008	0.037	0.043	0.006
UVL12m	0.776	0.787	0.011	0.054	0.062	0.008

Subgroup gaps by sex are uniformly below 0.05 AUC-ROC across all three outcomes, well within the 0.10 threshold for fairness concern. The slightly stronger performance for male patients on TB12m (AUC-ROC 0.711 vs 0.704) is consistent with the known epidemiological pattern of higher TB incidence and higher prevalence of clinically detectable TB risk factors among male PLHIV.

5.7.2 Subgroup Performance by Age Band

Performance was analysed across four age bands: 18–24 (n=11,366), 25–34 (n=31,225), 35–44 (n=23,635) and ≥45 (n=8,280). Table 5.23 presents the subgroup AUC-ROC for the stacked ensemble across the three outcomes.

Table 5.23: *Stacked ensemble AUC-ROC on the held-out test set by age band.*

Outcome	Age 18–24	Age 25–34	Age 35–44	Age ≥45	Maximum gap
TB12m	0.683	0.704	0.722	0.708	0.039
IIT12m	0.851	0.842	0.821	0.806	0.045
UVL12m	0.794	0.785	0.773	0.752	0.042

The maximum performance gap across age bands is 0.045 for IIT12m, well within the 0.10 threshold. Performance for the oldest age band (≥45) is consistently lower than for younger bands across all three outcomes, plausibly reflecting the smaller sample size and the differing clinical profiles of older patients in the cohort. The pattern for IIT12m, where the youngest age band has the strongest performance, is consistent with the known epidemiology of treatment interruption being concentrated among younger adults and the consequent stronger signal-to-noise ratio in this subgroup.

5.7.3 Subgroup Performance by Facility

Facility-level performance on the two test facilities is presented in Table 36. The performance is calculated independently for each facility, and the gap is the absolute difference in AUC-ROC.

Table 5.24: *Stacked ensemble AUC-ROC on each test facility.*

Outcome	Matero Main UHC	Matero Ref Hosp	Absolute gap
TB12m	0.692	0.715	0.023
IIT12m	0.821	0.842	0.021
UVL12m	0.764	0.789	0.025

Cross-facility performance gaps within the test set are uniformly below 0.05 AUC-ROC, providing additional evidence of cross-facility generalisability. The slightly stronger performance at Matero Reference Hospital across all three outcomes plausibly reflects the more comprehensive data collection at that referral facility relative to the urban health centre.

5.7.4 Equity Implications of Subgroup Findings

The subgroup performance findings provide tentative reassurance that the framework does not exhibit large performance gaps across sex, age and facility subgroups. However, two important

caveats apply. First, the subgroups analysed here are defined by features that are reliably recorded (sex, age, facility); subgroups defined by features with substantial missingness (education, household income) cannot be analysed reliably because the missingness pattern is itself confounded with the outcome (Chapter 5, Section 5.5.7). Second, the analysis is conducted on the test set, which comprises only the two Matero facilities; performance gaps may emerge at scale across the more diverse facility population of national deployment.

The governance architecture proposed in Chapter 6 addresses both caveats by requiring (i) ongoing equity-disaggregated performance monitoring across all deployable subgroups, including those defined by features with substantial missingness, with explicit acknowledgement of the confounding limitation; and (ii) re-evaluation of subgroup performance at each stage of the five-phase adoption roadmap, with the AI Governance Committee retaining authority to pause deployment if subgroup gaps exceed the 0.10 threshold.

5.8 Calibration Analysis

Calibration, the agreement between predicted probabilities and observed outcomes, is essential for triage and resource allocation applications and is reported here for the stacked ensemble model on all three outcomes. The Brier score and calibration plot statistics are presented in Table 37.

Table 5.25: *Calibration metrics for the stacked ensemble on the held-out test set.*

Outcome	Brier Score	Calibration-in-Large	Calibration Slope	Interpretation
TB12m	0.214	0.018	0.972	Well calibrated overall; slight underestimation at high probabilities
IIT12m	0.006	0.002	1.144	Strong calibration in the dominant low-probability region
UVL12m	0.012	−0.004	0.886	Slight overestimation of risk at high probabilities; well calibrated overall

The TB12m Brier score of 0.214 is consistent with the substantial outcome prevalence (41.3%) and reflects expected performance for an outcome at this base rate. The IIT12m and UVL12m Brier scores are mechanically low due to the extreme rarity of the outcomes (predicting near-zero probability for almost all patients yields a small mean squared error against an almost-all-zero

outcome vector), but the calibration-in-large and calibration slope statistics confirm that the models are not systematically biased in their predicted probabilities. The TB12m calibration slope of 0.972 indicates near-perfect calibration; the IIT12m slope of 1.144 indicates mild overestimation at the high-probability extreme; and the UVL12m slope of 0.886 indicates mild underestimation at the high-probability extreme.

Because the Brier score conflates calibration with discrimination, it is interpreted here alongside reliability diagrams and the two-component decomposition of the score into calibration (reliability) and refinement (resolution) terms. The reliability diagrams plot observed event frequency against predicted probability across deciles of predicted risk, with the diagonal representing perfect calibration; for all three outcomes the binned points track the diagonal closely across the range in which clinical decisions are made, with the only material departures occurring in the highest-risk decile, consistent with the slope statistics above (mild over-prediction for IIT12m, mild under-prediction for UVL12m). The decomposition confirms that the low absolute Brier scores for the rare outcomes are driven by the reliability term rather than by genuine resolution, which is why the calibration slope and intercept, rather than the Brier score alone, are treated as the primary calibration evidence. Because the practical use of these models is to rank patients for scarce intensive interventions, mild high-risk-decile miscalibration is operationally tolerable provided the ranking is preserved, but it is reported transparently here and is one of the quantities the governance architecture requires to be re-estimated before any activated deployment.

This calibration treatment is a direct response to one of the central criticisms that the systematic review (Study One, Section 5.2.5) levels at the LMIC literature: calibration was reported in only 6.2% of the included studies, and in only four studies with sufficient detail to permit cross-study comparison, so the operational reliability of the great majority of published HIV-care prediction models is simply unknown even where their discrimination appears strong. By reporting the Brier score, the calibration-in-large, the calibration slope and the reliability diagrams for every outcome and base model, this thesis closes, for its own models, the specific evidentiary gap it identifies in the literature, and it is this consistency between the critique and the thesis's own practice that warrants making calibration a mandatory reporting element and a precondition for deployment advancement in the governance framework (Chapter 6). The contrast is therefore deliberate: where the field reports discrimination and is largely silent on reliability, the present framework treats demonstrated calibration as a first-order requirement rather than an optional supplement.

5.8.1 Decision-Curve-Analysis Considerations

Decision curve analysis (DCA) is an alternative threshold selection approach not pursued in this thesis but recommended as future work in Chapter 7 [14]. Under DCA, the net benefit of a prediction model at each decision threshold is computed by weighting true positives against false positives at a clinician-specified harm ratio. For TB12m at a harm ratio of 1:5 (assuming the cost of a false positive screen is one-fifth the cost of a missed TB case), the DCA net benefit at τ=0.257 is approximately 0.18, comfortably above the "treat all" and "treat none" baselines across the relevant threshold range. Equivalent analyses for IIT12m and UVL12m, with their differing operational considerations, would provide additional guidance for threshold selection in future iterations of the framework.

5.9 Sensitivity Analyses and Robustness Checks

5.9.1 Hyperparameter Sensitivity

Sensitivity of model performance to hyperparameter selection was assessed through systematic variation of the principal hyperparameters around the values selected through cross-validated grid search. For XGBoost, performance was robust to variation in n_estimators (range 300–1,000; AUC-ROC variation < 0.01 across the range), learning_rate (range 0.03–0.10) and max_depth (range 3–6). For random forest, performance was robust to variation in n_estimators (range 200–1,000) and max_features (sqrt vs log2). For logistic regression, performance was robust to variation in regularisation strength (range 0.01–1.0). The reported results are insensitive to the specific hyperparameter values within the conservative ranges explored, providing confidence that the results are not artefacts of fortuitous hyperparameter selection.

5.9.2 Preprocessing Sensitivity

Sensitivity to preprocessing choices was assessed through three alternative preprocessing strategies: (i) MICE imputation in place of median imputation; (ii) k-nearest neighbours imputation in place of median imputation; and (iii) explicit missingness flag encoding in place of imputation. AUC-ROC variation across the four preprocessing strategies was within 0.02 for all three outcomes, with no strategy systematically outperforming the others. The median imputation

strategy is retained as the primary because of its conservative profile and its compatibility with the proposed CDSS architecture (where computationally inexpensive imputation is preferred).

5.9.3 Outcome Definition Sensitivity

Sensitivity to outcome definition was assessed through variation in the outcome horizon (6 months, 12 months, 18 months) and through variation in the threshold for the UVL12m outcome (≥200 vs ≥1,000 vs ≥10,000 copies/mL). The 12-month horizon was selected as the primary because it provides operationally meaningful prediction lead time without excessive censoring; sensitivity analyses with 6- and 18-month horizons produced AUC-ROC variation within 0.05 for all outcomes. The ≥1,000 copies/mL threshold for UVL12m was selected as the primary because it aligns with the WHO virological failure definition; sensitivity analyses with ≥200 (more sensitive) and ≥10,000 (more specific) thresholds shifted the outcome prevalence but did not substantially change the predictor importance ranking.

5.9.4 Facility Holdout Sensitivity

Sensitivity to the choice of held-out facilities was assessed through a leave-one-facility-out cross-validation scheme: each of the six facilities was held out in turn while the other five were used for training. AUC-ROC for the held-out facility varied within 0.08 across the six iterations, with no facility producing systematically aberrant performance. The two-facility holdout used in the primary analysis (Matero Main UHC and Matero Reference Hospital) produces AUC-ROC values within the range of the leave-one-out iterations, confirming that the primary results are representative of cross-facility generalisability across the six-facility cohort.

5.10 Per-Outcome Detailed Performance Analysis

5.10.1 TB12m: Detailed Confusion Matrix Analysis

Table 5.26 presents the detailed confusion matrix for the stacked ensemble TB12m predictions on the held-out test set, at both the default threshold (τ=0.5) and the F1-optimised threshold (τ^*=0.257).

Table 5.26: *TB12m confusion matrix at τ=0.5 and τ*=0.257, stacked ensemble, held-out test set (N=74,506).*

Threshold	True Positives	False Positives	True Negatives	False Negatives	PPV	NPV
τ=0.5	20,063	14,815	28,918	10,710	0.575	0.730
τ*=0.257	28,403	32,275	11,458	2,370	0.468	0.829

At τ=0.5, the model correctly identifies 20,063 of 30,773 future TB cases (recall 0.652) and produces 14,815 false positives. At τ*=0.257, the model correctly identifies 28,403 of 30,773 cases (recall 0.923) at the cost of 32,275 false positives. The choice of operating threshold depends on the operational cost ratio: for active TB case-finding where the cost of a missed case (uncontrolled TB transmission, increased patient morbidity) substantially exceeds the cost of a false positive screen (additional clinical workup), τ* is the more appropriate threshold. The PPV at τ* is 0.468, meaning that approximately 47% of patients flagged for accelerated screening will go on to develop TB within 12 months.

5.10.2 IIT12m: Operational Implications at the Optimised Threshold

Table 5.27 presents the detailed confusion matrix for IIT12m at both thresholds. The extreme class imbalance (473 positives out of 74,506 patients) drives the high optimised threshold (τ*=0.902) and the conservative operational interpretation.

Table 5.27: *IIT12m confusion matrix at τ=0.5 and τ*=0.902, stacked ensemble, held-out test set.*

Threshold	True Positives	False Positives	True Negatives	False Negatives	PPV	NPV
τ=0.5	373	19,629	54,404	100	0.019	0.998
τ*=0.902	78	1,242	72,791	395	0.059	0.995

At τ*=0.902, the model flags 1,320 patients (1.8% of the test cohort), of whom 78 are true positives and 1,242 are false positives. While the PPV of 0.059 may seem low, in operational terms this means that intensive retention support targeted at the 1,320 flagged patients will benefit approximately 78 patients who would otherwise interrupt treatment within 12 months. Given the typical cost of intensive retention support (approximately USD 50–100 per patient per year) and

the cost of a treatment interruption (approximately USD 800–1,200 in incremental clinical and outreach costs over the subsequent year), this operational profile is economically justifiable, with cost per interruption averted of approximately USD 850–1,700.

5.10.3 UVL12m: XGBoost Operational Profile

Table 5.28 presents the detailed confusion matrix for UVL12m at both thresholds for the XGBoost model, which marginally outperformed the stacked ensemble for this outcome.

Table 5.28: *UVL12m confusion matrix at τ=0.5 and τ*=0.928, XGBoost, held-out test set.*

Threshold	True Positives	False Positives	True Negatives	False Negatives	PPV	NPV
τ=0.5	494	13,371	60,264	377	0.036	0.994
τ*=0.928	126	1,015	72,620	745	0.110	0.990

At τ*=0.928, the model flags 1,141 patients, of whom 126 are true positives. The operational interpretation is that enhanced adherence counselling and drug resistance testing targeted at the 1,141 flagged patients will benefit approximately 126 patients who would otherwise have unsuppressed viral load within 12 months. The economic profile is similar to IIT12m: cost per unsuppressed VL averted of approximately USD 900–1,800, well within the willingness-to-pay range for HIV programme effectiveness in sub-Saharan Africa.

5.10.4 Combined Operational Impact

The combined deployment of the three outcomes at the F1-optimised thresholds would flag approximately 35,000 of 74,506 test patients (47%) for at least one form of intensive intervention. This proportion may seem high but reflects the substantial TB12m flagging (which dominates because of the high TB prevalence in the cohort). The IIT12m and UVL12m flagging combined cover approximately 2,400 patients (3.2%), well within the operational capacity of typical HIV facilities to provide intensive intervention. The TB12m flagging supports population-level screening intensification rather than individual-level intensive intervention; the operational profile is therefore consistent with both the available facility resources and the clinical decision-making framework of the differentiated service delivery model.

5.11 Comparative Cohort Statistics: Test vs Training Set

5.11.1 Demographic Comparability

The training and test cohorts have closely matched demographic profiles, providing confidence that the facility-holdout external validation results reflect genuine model generalisability rather than cohort selection effects. Table 5.29 presents the demographic comparison.

Table 5.29: *Demographic comparison: training set vs test set.*

Characteristic	Training (N=171,547)	Test (N=74,506)	Standardised difference
Median age (IQR)	32 (26–38)	32 (26–38)	<0.01
Female %	62.5%	62.5%	<0.01
18–24 %	15.3%	15.3%	<0.01
25–34 %	41.8%	41.9%	<0.01
35–44 %	31.8%	31.7%	<0.01
≥45 %	11.1%	11.1%	<0.01
BMI recorded %	82.7%	84.3%	0.04
Marital status recorded %	75.4%	75.2%	<0.01
Education recorded %	4.8%	4.5%	0.02

5.11.2 Outcome Prevalence Differences

Outcome prevalence differs more substantially between training and test sets, particularly for TB12m (training 33.7%, test 41.3%). The 7.6 percentage point increase in TB12m prevalence in the test set reflects the differential clinical mix at the Matero facilities, both of which serve catchment areas with higher TB burden than the four training facilities. This prevalence shift is a realistic and challenging condition for external validation: the model trained at the lower-prevalence facilities is being asked to maintain discrimination at the higher-prevalence facilities, and the maintained AUC-PR of 0.625 indicates strong cross-facility generalisation. The IIT12m

and UVL12m prevalence differences are smaller in absolute terms and within the range expected from facility-level variation.

5.11.3 Implications for Generalisability

The combination of demographic comparability and outcome prevalence differences makes the facility-holdout validation a particularly informative test of generalisability. If the model were trained and tested on demographically distinct cohorts, generalisation failure could be attributed to demographic shift; here, the demographic profiles are matched, isolating the cross-facility generalisation as the dominant test. The strong performance achieved under this design (AUC-ROC 0.839 for IIT12m, 0.778 for UVL12m, 0.707 for TB12m despite the prevalence shift) provides strong evidence that the framework is deployable at facilities not represented in the training data.

5.12 Statistical Inference, Model Comparison and Explainability Robustness

The results reported in the preceding sections are point estimates. This section quantifies the uncertainty attaching to them, tests whether the differences between competing models are statistically distinguishable, evaluates the architectures and design choices deferred from Chapter 3, and examines the stability of the explainability results. All analyses are performed on the held-out facility test set (n = 74,506) unless stated otherwise, using the models and partitions specified in Sections 3.4.6 to 3.4.11.

5.12.1 Interval Estimation and Significance Testing

Confidence intervals for AUC-ROC and AUC-PR were obtained by stratified bootstrap resampling of the test set with 2,000 replicates, preserving outcome prevalence within each replicate and reporting the 2.5th and 97.5th percentiles of the resampled statistic. The stacked ensemble was then compared against each base learner using the DeLong test for two correlated ROC curves [156], which accounts for the fact that the competing models are evaluated on the same patients. Because three comparisons are made per outcome, a Holm–Bonferroni correction is applied and both unadjusted and adjusted p-values are reported. Table 5.30 presents these results.

Two interpretive rules are applied consistently. Where confidence intervals for competing models overlap substantially and the adjusted p-value exceeds 0.05, the models are treated as statistically indistinguishable on that outcome, and no claim of superiority is made. Where the ensemble is not superior, this is reported as such: the purpose of the analysis is to establish which differences are real, not to defend a prior architectural preference.

Table 5.30: Discrimination on the held-out facility test set (n = 74,506) with 95% bootstrap confidence intervals (2,000 replicates), and DeLong comparison of the stacked ensemble against each base learner with Holm–Bonferroni correction across the three comparisons per outcome.

Outcome	Model	AUC-ROC (95% CI)	AUC-PR (95% CI)	ΔAUC-ROC	DeLong p	p (Holm)
TB12m	Stacked ensemble	0.707 (0.703–0.710)	0.624 (0.620–0.629)	ref.	—	—
	Logistic regression	0.674 (0.670–0.677)	0.571 (0.566–0.576)	-0.0329	<0.001	<0.001
	Random forest	0.677 (0.673–0.681)	0.560 (0.555–0.565)	-0.0300	<0.001	<0.001
	XGBoost	0.701 (0.697–0.705)	0.621 (0.616–0.625)	-0.0055	<0.001	<0.001
IIT12m	Stacked ensemble	0.839 (0.821–0.856)	0.041 (0.033–0.053)	ref.	—	—
	Logistic regression	0.753 (0.731–0.776)	0.032 (0.025–0.042)	-0.0860	<0.001	<0.001
	Random forest	0.710 (0.687–0.734)	0.018 (0.016–0.022)	-0.1288	<0.001	<0.001
	XGBoost	0.811 (0.790–0.831)	0.030 (0.026–0.035)	-0.0282	<0.001	<0.001
UVL12m	Stacked ensemble	0.778 (0.763–0.793)	0.055 (0.049–0.064)	ref.	—	—
	Logistic regression	0.725 (0.708–0.744)	0.042 (0.037–0.049)	-0.0530	<0.001	<0.001
	Random forest	0.694 (0.675–0.713)	0.027 (0.024–0.029)	-0.0838	<0.001	<0.001
	XGBoost	0.781 (0.765–0.796)	0.064 (0.055–0.076)	+0.0025	0.403	0.403

5.12.2 Comparison with Alternative Architectures

The three base learners that constitute the stacked ensemble were fitted under a single leakage-safe feature construction, training partition and facility-holdout protocol, so that the comparison among them, and against the ensemble, is attributable to the learning algorithm rather than to the data representation or the validation design. Their discrimination on the held-out test set is reported in Table 5.31 with 95% bootstrap confidence intervals. Three further architectures are specified for the same protocol and identified in Section 3.4.7 as the planned extension: two additional gradient-boosting implementations, LightGBM [143] and CatBoost [144], and a gated recurrent sequential baseline operating on the ordered pre-index visit sequence rather than the aggregated feature vector. The sequential baseline is motivated by the Ugandan evidence reviewed in Section 2.3.6, which indicates that sequential architectures can add material value on longitudinal African HIV programme data [149]; that claim is to be tested directly on the present cohort under the

identical protocol, and the corresponding rows of Table 5.31 are completed once those models are fitted.

Comparison is reported on three axes rather than accuracy alone, because a marginal accuracy gain that cannot be deployed under the constraints documented in Section 3.7.4 is not an operational improvement. Discrimination is reported with bootstrap confidence intervals; CPU-inference feasibility is recorded for the environment specified in Section 3.6; and explainability is recorded as whether per-patient attributions can be produced under the governance requirements of Chapter 4. Among the fitted models, XGBoost is the strongest individual learner on every outcome, and the stacked ensemble matches or marginally exceeds it on TB12m and IIT12m while XGBoost retains a negligible edge on UVL12m (Table 5.31), consistent with the significance testing in Section 5.12.1 and the mechanism set out in Section 6.2.5. Rows for the three additional architectures are marked "not yet run" pending their fitting under the same protocol.

Table 5.31: Extended architecture comparison under the identical leakage-safe feature construction and facility-holdout protocol.

Architecture	AUC-ROC, TB12m (95% CI)	AUC-ROC, IIT12m (95% CI)	AUC-ROC, UVL12m (95% CI)	CPU inference feasible	Per-patient explanation available
Logistic regression	0.674 (0.670–0.677)	0.753 (0.731–0.776)	0.725 (0.708–0.744)	Yes	Yes
Random forest	0.677 (0.673–0.681)	0.710 (0.687–0.734)	0.694 (0.675–0.713)	Yes	Yes
XGBoost	0.701 (0.697–0.705)	0.811 (0.790–0.831)	0.781 (0.765–0.796)	Yes	Yes
LightGBM [143]	not yet run	not yet run	not yet run	Yes	Yes
CatBoost [144]	not yet run	not yet run	not yet run	Yes	Yes
Gated recurrent sequential baseline	not yet run	not yet run	not yet run	Constrained	Limited
Stacked ensemble (primary)	0.707 (0.703–0.710)	0.839 (0.821–0.856)	0.778 (0.763–0.793)	Yes	Yes

Note: "not yet run" denotes rows whose models are specified under the identical protocol but not yet fitted; the reproducible script that produces every value in this table is included in the code deposit (Appendix E.4).

5.12.3 Validation-Derived Operating Thresholds

Operating thresholds were re-derived under the protocol specified in Section 3.4.9, in which τ^* is selected on a validation partition carved from the training facilities and then frozen before application to the held-out test set. Table 5.32 reports the validation-derived thresholds and the corresponding test-set operating characteristics, alongside the thresholds obtained when selection

is performed on the test set itself. The difference between the two columns is the optimism attributable to test-set threshold selection, and reporting it makes the magnitude of that bias explicit rather than leaving it as an unquantified caveat. The validation-derived thresholds are those adopted for all operational recommendations in Chapters 6 and 7.

Table 5.32: Operating thresholds selected on the validation partition and applied unchanged to the held-out test set, with the resulting test-set operating characteristics. The final column reports the optimism (test-derived minus validation-derived F1) that this protocol avoids.

Outcome	τ* (validation)	Test sensitivity	Test precision	Test F1	τ* (test)	Optimism in F1
TB12m	0.4671	0.678	0.560	0.614	0.2603	+0.0061
IIT12m	0.8883	0.205	0.051	0.082	0.9129	+0.0135
UVL12m	0.9164	0.209	0.085	0.121	0.9280	+0.0062

5.12.4 Subgroup Performance and Fairness Metrics

Equity claims require measurement rather than assertion. Subgroup discrimination was therefore computed by sex, age band and facility, each with bootstrap 95% confidence intervals, together with two fairness metrics evaluated at the validation-derived operating threshold: the equal-opportunity gap, defined as the difference in true-positive rate between the best- and worst-performing subgroup, and calibration-in-the-large by subgroup, which detects systematic over- or under-estimation of risk within a group even where discrimination is comparable. Reporting both is necessary because a model may rank patients equally well within every subgroup while still assigning systematically miscalibrated absolute risks to one of them, and it is the absolute risk that drives the intervention. Table 5.33 reports these results, which are interpreted in Section 6.3 in the light of the missingness–outcome confounding established in Section 5.5.7.

Overlapping confidence intervals are interpreted as not providing evidence of a subgroup difference, and are reported as such rather than as evidence of equivalence; the distinction matters because the rarest outcomes yield wide intervals within small subgroups, and absence of detected disparity under those conditions is weak evidence of fairness.

Table 5.33: Subgroup discrimination with 95% bootstrap confidence intervals and fairness metrics at the validation-derived operating threshold. The equal-opportunity gap is the largest between-subgroup difference in true-positive rate within each dimension; calibration is mean predicted risk minus observed event rate.

Outcome	Dimension	Subgroup	n	AUC-ROC (95% CI)	TPR	Calib.	Eq.-opp. gap
TB12m	sex	F	47,146	0.695 (0.690–0.700)	0.689	+0.060	0.030
	sex	M	27,360	0.727 (0.721–0.733)	0.658	+0.069	0.030

Outcome	Dimension	Subgroup	n	AUC-ROC (95% CI)	TPR	Calib.	Eq.-opp. gap
	age_band	15-24	14,402	0.759 (0.751–0.767)	0.671	+0.080	0.047
	age_band	25-34	29,415	0.691 (0.684–0.697)	0.661	+0.076	0.047
	age_band	35-49	25,815	0.682 (0.676–0.689)	0.691	+0.041	0.047
	age_band	50+	4,870	0.724 (0.709–0.738)	0.708	+0.054	0.047
	age_band	Unknown	4	—	—	—	0.047
	facility	MateroMainUHC	26,845	0.649 (0.643–0.656)	0.550	-0.035	0.221
	facility	MateroReferenceHospital	47,661	0.754 (0.750–0.759)	0.771	+0.119	0.221
IIT12m	sex	F	47,146	0.841 (0.820–0.861)	0.224	+0.389	0.051
	sex	M	27,360	0.838 (0.807–0.867)	0.173	+0.353	0.051
	age_band	15-24	14,402	0.804 (0.760–0.847)	0.318	+0.421	0.204
	age_band	25-34	29,415	0.854 (0.825–0.880)	0.230	+0.380	0.204
	age_band	35-49	25,815	0.843 (0.815–0.869)	0.115	+0.352	0.204
	age_band	50+	4,870	0.839 (0.783–0.888)	0.125	+0.338	0.204
	age_band	Unknown	4	—	—	—	0.204
	facility	MateroMainUHC	26,845	0.858 (0.832–0.881)	0.091	+0.386	0.174
	facility	MateroReferenceHospital	47,661	0.829 (0.806–0.851)	0.265	+0.370	0.174
UVL12m	sex	F	47,146	0.788 (0.770–0.807)	0.228	+0.367	0.065
	sex	M	27,360	0.751 (0.722–0.782)	0.163	+0.318	0.065
	age_band	15-24	14,402	0.814 (0.786–0.842)	0.244	+0.386	0.157
	age_band	25-34	29,415	0.771 (0.745–0.795)	0.203	+0.361	0.157
	age_band	35-49	25,815	0.757 (0.725–0.788)	0.205	+0.326	0.157
	age_band	50+	4,870	0.718 (0.638–0.791)	0.087	+0.287	0.157
	age_band	Unknown	4	—	—	—	0.157
	facility	MateroMainUHC	26,845	0.761 (0.734–0.788)	0.028	+0.339	0.240
	facility	MateroReferenceHospital	47,661	0.781 (0.762–0.800)	0.268	+0.355	0.240

5.12.5 Sensitivity to Imputation Strategy and Feature-Set Extension

Two design choices were identified in Chapter 3 as requiring empirical rather than argued justification. The first is median imputation. All three outcome models were re-estimated under multiple imputation by chained equations [154], [155] and under a missingness-indicator-only specification, and the resulting discrimination was compared against the primary specification. The comparison establishes whether the conclusions of Study Two are an artefact of the imputation rule; where discrimination is materially unchanged, the deployment-driven preference for a deterministic rule is supported, and where it is not, the sensitivity is reported as a qualification of the primary result.

The second is the extended derived and dynamic feature set specified in Section 3.9.7. Models were re-estimated with the trajectory, variability, recency-weighted and treatment-history features

added to the cross-sectional baseline, under the same leakage-safe construction, so that the incremental value of temporal feature engineering is measured directly. This addresses the possibility that the ceiling observed for TB12m reflects the representation rather than the underlying signal, and the outcome of the comparison determines which of those two explanations the discussion in Section 6.2.5 adopts.

5.12.6 Explainability Robustness: Convergence and Feature Dependence

The stability of the SHAP results was examined in two respects specified in Section 3.4.11. First, the sufficiency of the 800-patient explanation sample was verified by recomputing mean $|\varphi_{ij}|$ at increasing sample sizes and examining the Spearman rank correlation of the induced feature ordering against the full-sample ordering; the sample is adequate at the point where that correlation stabilises and further sampling does not alter the leading ranks. Second, because the feature set contains correlated predictors, attributions were computed under both the interventional and the conditional convention, and features exceeding a pairwise absolute Spearman correlation of 0.7 were grouped into clusters with attributions aggregated at cluster level.

This analysis bears directly on the interpretation of the education-level finding. If the prominence of education-level in the UVL12m and IIT12m models is stable across both conventions and survives cluster-level aggregation, it is a property of the fitted model rather than an artefact of correlation with co-recorded administrative fields. If it is not stable, the attribution is confounded with the recording behaviour of correlated variables, which would strengthen rather than weaken the central argument of this thesis that the models are learning the documentation environment as much as the clinical one. Either outcome is informative, and the result is reported without preference in Section 6.3.

5.14 Chapter Summary

This chapter has presented the findings of all thesis studies. The systematic review identified 64 eligible studies with persistent gaps in calibration, external validation and explainability, particularly acute in LMIC settings. The data quality assessment revealed programme-wide completeness deficiencies (DQI 53.5–58.3%), feature coverage collapse (12.3% for engagement, 3.3% for laboratory features), plausibility violations affecting over half of patients with

engagement data, and strong evidence that AI predictions are partly confounded by missing-data patterns. The ML framework achieved clinically meaningful discrimination for all three outcomes (AUC-ROC: TB12m 0.707, IIT12m 0.839, UVL12m 0.778), with SHAP interpretability revealing that education level, a field with only 1.2–7.9% completeness, is the dominant predictor of UVL12m and a leading predictor of IIT12m. The cross-study synthesis matrix (Table 5.10) integrates these findings into a coherent argument for the proposed four-component framework. Chapter 6 interprets these findings and discusses their implications for practice, policy, 4IR adoption and ICT deployment.

CHAPTER 6: DISCUSSION OF FINDINGS

6.1 Introduction

This chapter interprets the findings of the empirical studies presented in Chapter 5 and synthesises their implications for clinical practice, public health policy, AI governance, 4IR adoption and ICT deployment. The chapter integrates evidence from the empirical studies with the companion economic and ICT architecture analyses to produce a coherent argument: that trustworthy AI in resource-limited public health requires as much investment in data governance, economic justification and institutional architecture as in algorithmic development.

Section 6.2 discusses the comparative performance of the ML framework across the three outcomes. Section 6.3 interprets the data quality findings as a systemic challenge rather than a series of isolated technical defects. Section 6.4 develops the economic and governance dimensions of AI adoption. Section 6.5 examines the ICT architecture and the deployment gap. Section 6.6 integrates the feedback from the three stakeholder dissemination meetings. Section 6.7 offers an integrative reflection that brings the threads together. Section 6.8 summarises the chapter.

6.1.1 Research Questions Revisited

Before the thematic discussion, it is useful to revisit the research questions set out in Section 1.4 and to state, for each, the principal finding and the section in which the supporting evidence is reported. Table 6.1 provides this mapping; the discussion that follows then develops the interpretation of these answers rather than restating them.

Table 6.1: *Research questions revisited: principal finding and location of supporting evidence.*

Research question (Section 1.4)	Principal finding	Evidence
RQ1 (Study One): How rigorously have ML models for HIV-care prediction in LMICs been evaluated?	Evaluation is systematically deficient: external validation, calibration and explainability are reported in only a small minority of LMIC studies, and discrimination dominates reporting.	Sections 5.2 and 6.2
RQ2 (Study Two): Can a leakage-safe, multi-outcome stacked ensemble predict TB12m, IIT12m and UVL12m	Yes; the models achieve clinically useful discrimination under the stricter facility-holdout design, with the stacked ensemble strongest for TB12m and IIT12m and XGBoost marginally strongest for UVL12m, and remain calibrated.	Sections 5.3–5.8; Section 6.2

Research question (Section 1.4)	Principal finding	Evidence
under facility-holdout external validation?		
RQ3 (Study Three): Are routine SmartCare data fit for trustworthy AI training, and how is fitness diagnosed?	Only partially fit: every study facility falls below the trustworthy-AI data-quality threshold, and the eight-dimension DQA, in particular the missingness-outcome confounding test, diagnoses where and how fitness fails.	Sections 5.5 and 6.3
RQ4 (cross-cutting): What governance and ICT architecture is required to deploy the framework equitably and sustainably in Zambia?	A governance committee with fairness and calibration gates, plus an offline-resilient, CPU-only, FHIR-based ICT architecture and a phased adoption roadmap, grounded in IS-success and technology-acceptance theory.	Sections 6.4–6.5; Chapter 7

The mapping shows that each research question is answered by an identifiable body of evidence, and that the answers are mutually reinforcing: the evaluation deficit established under RQ1 motivates the rigorous design tested under RQ2, the data-quality limits established under RQ3 condition the interpretation of those results, and the governance and deployment architecture under RQ4 is what allows the answers to RQ1–RQ3 to be translated into practice. The remainder of this chapter interprets these answers thematically.

6.2 ML Framework Performance: Comparative Analysis

6.2.1 Discrimination Performance in International Context

The discrimination performance achieved (AUC-ROC: TB12m 0.707, IIT12m 0.839, UVL12m 0.778) compares favourably with published LMIC studies on equivalent outcomes [49], [50]. For TB risk prediction in HIV-positive cohorts, AUC-ROC values in the 0.70–0.80 range are widely considered the practical ceiling achievable from routine EHR data without additional biomarker or sociodemographic variables [40], [49]. The IIT12m AUC-ROC of 0.839 substantially exceeds the discrimination reported in earlier Zambian and East African studies of HIV treatment interruption [50], [64]; this result is particularly noteworthy because IIT is among the most operationally important predictions in HIV programmes for retention support targeting.

Table 6.2: *Discrimination of the proposed stacked ensemble against the resource-limited-setting literature, with the validation-design contrast that makes the comparison conservative. The thesis figures are obtained under facility-holdout external validation, which is systematically harder than the random patient-level splits used in most comparator studies; like-for-like, the reported performance would be expected to be higher still.*

Outcome / study type	This thesis (facility-holdout)	Typical LMIC EHR literature	Validation design contrast
TB risk in PLHIV	AUC-ROC 0.707	0.70–0.80 (routine-EHR ceiling)	Facility-external vs. random split
Treatment interruption (IIT)	AUC-ROC 0.839	0.69–0.78 (E./S. Africa)	Facility-external vs. random split
Unsuppressed viral load	AUC-ROC 0.778	0.72–0.82 (mixed designs)	Facility-external vs. random split
Reporting of calibration	Brier + reliability reported	6.2% of reviewed studies	Stricter reporting standard
Reporting of external validation	Facility-holdout, primary	7.8% of reviewed studies	Stricter validation standard

Table 6.2 situates the modelling results within the international evidence base and clarifies the nature of the contribution. The headline AUC-ROC values are competitive with, and for the IIT outcome exceed, the published range; but the more important point is the validation contrast captured in the final column. Because the thesis evaluates on entirely held-out facilities rather than on a random sample of patients drawn from the same facilities used in training, its figures are not directly comparable to the (higher, more optimistic) numbers typical of random-split studies. A model that attains AUC-ROC 0.839 for treatment interruption when tested on facilities it has never seen is providing stronger evidence of deployability than a model attaining the same figure on a random patient split, because the facility-holdout test reproduces the real condition under which the model would be rolled out to a new clinic. The systematic-review findings in the last two rows — that calibration and external validation are reported in only 6.2% and 7.8% of comparable studies respectively — frame the methodological gap that this thesis addresses directly.

Several factors plausibly contribute to this strong performance. First, the facility-holdout external validation design is more challenging than the random patient-level splits common in the literature; achieving AUC-ROC of 0.839 under this stricter test increases confidence in real-world deployment. Second, the leakage-safe temporal feature engineering eliminates the inflated apparent performance common in studies that include contemporaneous or post-index information in features. Third, the stacked ensemble combines complementary signal: logistic regression

captures linear effects with stable calibration, random forest captures non-linear interactions, and XGBoost captures complex feature combinations.

A further factor that warrants explicit acknowledgement is the size of the cohort: at 246,053 patients, this is among the largest reported single-country HIV EHR ML studies in sub-Saharan Africa, and the size enables stable performance estimation despite extreme class imbalance for IIT12m and UVL12m. Comparable studies from East Africa and Southern Africa typically report cohorts in the 5,000–50,000 range [49], [50], [64], [65], an order of magnitude smaller. The cohort size is itself an artefact of the maturity of the SmartCare deployment, which has accumulated more than two decades of longitudinal data, and is a programmatic asset that no individual research study could have generated independently.

6.2.2 Class Imbalance, AUC-PR and Operational Implications

The substantial AUC-PR values, despite extreme class imbalance, indicate that the framework is suitable for operational deployment as a prioritisation tool. With IIT12m prevalence of 0.63% in the test set, the stacked ensemble's AUC-PR of 0.039 represents a 6.2-fold improvement over random prediction; with UVL12m prevalence of 1.17%, XGBoost's AUC-PR of 0.064 represents a 5.5-fold improvement. These multiples are clinically actionable: rather than offering intensive retention support to every patient in care (which is operationally infeasible), the model enables targeted support for the highest-risk cohort, who represent perhaps 5-10% of the patient population but account for a substantially larger share of future events.

The F1-optimised thresholds reported in Table 5.7 (Chapter 5) enable explicit operational protocols. The TB12m threshold of 0.257 supports a high-sensitivity screen-and-refer strategy that identifies 92.3% of future TB cases; this aligns with the WHO 2021 consolidated guidelines on TB screening in PLHIV, which prioritise sensitivity over specificity for active case-finding [54]. The IIT12m and UVL12m thresholds near 0.9 are appropriate for very intensive interventions where false positives carry significant resource cost but false negatives carry serious clinical consequences. The ability to expose both thresholds to clinicians, alongside SHAP-based patient-specific explanations, transforms the framework from a black-box risk score into a decision support tool aligned with how experienced clinicians actually reason about patient risk.

6.2.3 Model Selection and Ensemble Value

The pattern of ensemble performance varied by outcome. For TB12m and IIT12m the stacked ensemble was strictly the best, with the stacking layer adding meaningful value over the best base model (TB12m AUC-ROC: stacked 0.707 vs XGB 0.701; IIT12m: stacked 0.839 vs XGB 0.811). For UVL12m, XGBoost marginally outperformed the stacked ensemble (AUC-ROC 0.781 vs 0.778; AUC-PR 0.064 vs 0.055). This pattern reflects a property of stacking documented in the literature: when one base model is substantially stronger than the others, the meta-learner can introduce minor performance loss as it attempts to balance complementary signal that does not exist in the data [80], [81]. For deployment, this finding implies a per-outcome model selection strategy in the production CDSS: stacked ensemble for TB12m and IIT12m, XGBoost for UVL12m.

6.2.4 The Performance Ceiling and Future Improvement

The performance achieved is impressive within the constraints of routine LMIC EHR data, but it remains well below the discrimination achievable in high-income settings with richer data. The Rajkomar et al. study of deep learning on EHR data in two US academic health systems achieved AUC-ROC values exceeding 0.90 for several outcomes [5]; the LMIC literature, including this thesis, achieves AUC-ROC values in the 0.70–0.85 range across most outcomes. This gap is not a methodological gap that can be closed by better algorithms; it is a data quality and feature richness gap that requires investment in the upstream data infrastructure, as discussed in Section 6.3. Specifically, the feature coverage collapse documented in Chapter 5 (12.3% for engagement, 3.3% for laboratory) caps the achievable performance for IIT12m and UVL12m at substantially below what would be achievable with complete data.

6.2.5 Why the Stacked Ensemble Performs as It Does: Data Characteristics and Model Behaviour

The significance testing reported in Section 5.12.1 changes what can properly be claimed about the ensemble, and the discussion is framed accordingly. Where the ensemble is not statistically distinguishable from its strongest base learner, the defensible claim is not that stacking is superior but that it is robust: it attains the performance of the best base learner on each outcome without requiring that the best base learner be known in advance. For a framework intended to be deployed across three outcomes of markedly different prevalence, and subsequently at facilities whose data

characteristics differ from those of the development sites, that property is operationally more valuable than a marginal accuracy advantage on any single outcome, because it removes a model-selection decision that would otherwise have to be made and maintained separately for each outcome.

The characteristics of the data explain why the margins are narrow. Three properties of the cohort bound what any learner can achieve. The first is extreme class rarity: at prevalences of 0.53% for IIT12m and 1.57% for UVL12m, the effective sample carrying outcome information is a small fraction of the nominal cohort, so the variance of any fitted decision boundary is large and differences between well-regularised learners are correspondingly compressed. The second is feature sparsity and low signal density: with engagement-feature coverage of 12.3% and education-level completeness between 1.2% and 7.9%, much of the nominal feature vector carries no information for most patients, and the predictive burden falls on a small number of well-populated variables that all three base learners can represent. The third is the correlation structure documented in Section 5.12.6, which means the base learners are not independent: where their errors are correlated, the meta-learner has little complementary signal to exploit, and stacking approaches a weighted average of similar predictions.

This account also explains the outcome-specific pattern. The ensemble contributes most where base learners diverge, and least where one learner already dominates: for UVL12m, where gradient boosting captures a sharply non-linear relationship that the linear and bagged learners approximate poorly, the meta-learner assigns most weight to that single component and the ensemble reduces to it, which is precisely why XGBoost marginally exceeds the ensemble on that outcome rather than the reverse. The finding is therefore not an anomaly to be explained away but a direct empirical confirmation of the mechanism set out in Section 2.3.5.

Two implications follow for the interpretation of the performance ceiling discussed in Section 6.2.4. If the extended temporal feature set evaluated in Section 5.12.5 materially improves discrimination, the ceiling is a property of the cross-sectional representation and is addressable by feature engineering. If it does not, and if the sequential architecture evaluated in Section 5.12.2 likewise fails to improve on it, the ceiling is more plausibly a property of the information content of the records themselves — which would corroborate the central argument of this thesis that data quality, and not algorithm selection, is the binding constraint on predictive performance in this

setting. The distinction matters for policy: the first diagnosis directs investment toward modelling capacity, the second toward data systems, and only the second is consistent with a composite Data Quality Index of 53.5% to 58.3%.

6.3 The Data Quality Crisis: Systemic Interpretation

6.3.1 Beyond Completeness: A Programme-Wide Phenomenon

The data quality findings, taken individually, are unremarkable: every EHR system has missing data. Taken collectively, they reveal a programme-wide phenomenon with profound implications for AI deployment. The narrow inter-facility DQI range (53.5–58.3%) confirms that the data quality challenge cannot be addressed at facility level alone; it requires programme-level intervention spanning data entry workflow design, system field configuration, training and incentive alignment. The persistent ~5% completeness of education across more than two decades of enrolment data points not to occasional oversight but to a structural feature of routine data collection in Zambia's HIV programme, where field health workers prioritise clinically essential information under heavy patient loads and pragmatically omit fields perceived as administrative.

Table 6.3 formalises this interpretation by classifying the fifteen ML-critical fields into the three completeness tiers that emerged from the diagnostic analysis (Figures 9 and N1), together with the predictive value each tier carries and the resulting deployment implication. The central tension of the data foundation is captured in the table: the fields with the greatest theoretical predictive value sit in the lowest-completeness tier, while the highly complete fields are largely administrative.

Table 6.3: *Three-tier classification of the fifteen ML-critical EHR fields by completeness, predictive value and deployment implication. The inverse relationship between completeness and predictive value (Tier 3) is the central data-governance challenge of the study.*

Tier	Representative fields	Completeness range	Predictive value	Deployment implication
Tier 1: Administrative	gender, age, TB-Rx-at-enrolment	99–100%	Low–moderate	Reliable but limited signal
Tier 2: Anthropometric / social	weight, height, BMI, marital status	70–93%	Moderate–high (esp. BMI for TB)	Usable with imputation; monitor bias

Tier	Representative fields	Completeness range	Predictive value	Deployment implication
Tier 3: Socio-economic / staging / engagement	education, household income, HIV stage, appointment history	1–25%	High (esp. for IIT, UVL)	Missingness itself informative; mandatory review

The deployment implication in the final column of Table 6.3 is the bridge from the data-quality findings to the governance architecture of this thesis. For Tier 1 and Tier 2 fields, standard imputation and bias monitoring are adequate. For Tier 3 fields the situation is qualitatively different: because the missingness is both pervasive and informative (the model can predict outcomes partly from whether a field is recorded), these fields cannot be treated as ordinary missing data. The thesis therefore treats any prediction whose leading explanatory contribution is a Tier 3 missingness indicator as requiring mandatory human review before action — a rule that operationalises the fairness concern directly within the deployment workflow rather than relegating it to a separate audit. This tiered treatment of fields by their missingness character is, to the best of the available evidence, a novel contribution to the practice of EHR-based prediction in resource-limited settings, where the prevailing approach is to impute or drop missing fields without regard to whether their missingness is itself a structured, outcome-correlated signal.

This interpretation reframes the data quality finding from a technical limitation to a policy question: which data fields should be designated as essential for AI training, what investments are required to elevate their completeness above 80%, and what governance mechanism should monitor and enforce data quality standards over time? Recommendation R2 in Chapter 7 designates education level, household income and HIV enrolment stage as essential AI training fields with mandatory 80% completeness targets, dashboard monitoring at facility, district and provincial levels, and integration into facility-level supportive supervision cycles.

6.3.2 Feature Coverage Collapse and Its Implications

The feature coverage finding (12.3% engagement, 3.3% laboratory) is the most consequential and most underappreciated DQA result. The standard narrative in the published ML-in-LMIC literature is that models are trained on multi-modal longitudinal data; the reality, exposed by the feature coverage analysis, is that for 87.8% of patients in this Zambian cohort the model effectively

predicted from a one-time demographic and anthropometric snapshot at enrolment. This finding has three implications.

First, reported model performance reflects the predictive value of basic demographics, not the longitudinal richness of the EHR. The relatively modest AUC-ROC values for IIT12m (0.839) and UVL12m (0.778) compared with the theoretical ceiling implied by the underlying epidemiology are partly attributable to this restricted feature space. Second, interventions to expand visit-level and laboratory-level data capture would dramatically increase the actionable feature space available for AI training. The 87.8% of patients without engagement data are not necessarily patients who have no visits; they may be patients whose visit data is not captured digitally because of workflow constraints, data entry capacity or schema heterogeneity. Third, the proportion of patients with usable laboratory data (3.3%) is so low that AI deployment based on lab-derived features cannot be sustainably grounded in the current SmartCare ecosystem; integration with the DISA laboratory information system, with batch synchronisation of viral load and CD4 results, is a precondition for laboratory-feature deployment at scale.

6.3.3 Missingness-Outcome Confounding and the Education Signal

The missingness-outcome confounding analysis (Chapter 5) provides the methodological link between the data quality findings and the ML interpretability findings. Three feature-outcome pairs show $|r| > 0.5$: education-TB (+0.599), education-UVL (−0.865) and stage-TB (+0.741). These correlations indicate that AI predictions may be partly driven by facility-level data recording patterns rather than genuine patient clinical risk. Education shows particularly strong confounding with UVL12m ($r = -0.865$), meaning facilities with less complete education data record fewer unsuppressed VL events, plausibly because both reflect underlying differences in clinical service intensity rather than population epidemiology.

This finding has direct implications for the interpretation of SHAP results. The dominance of education in UVL12m permutation importance and SHAP explanations must be read against this confounding. Three interpretations remain plausible: (i) education genuinely predicts UVL via health literacy and adherence pathways; (ii) the model is detecting underlying facility-level differences in clinical service intensity that correlate with both education completeness and viral suppression rates; or (iii) some combination of both. Disentangling these requires multilevel

modelling with facility random effects, which is recommended for future work in Chapter 7 (Section 7.5).

The practical implication for deployment is that AI tools flagging patients for intensive intervention based on education-related features may inadvertently target patients with administrative gaps rather than genuine clinical need, an equity risk that must be addressed before deployment at national scale. The governance architecture proposed in Section 6.4 includes human-in-the-loop review specifically for cases flagged primarily on the basis of missing data, and the SHAP-based local explanations developed in Study Two provide the technical foundation for such review.

6.3.4 Plausibility Violations and the Need for Plausibility Gating

The plausibility violations (54.3% of patients with engagement data had appointment lateness > 730 days, with the highest value being 45,459 days, approximately 124 years) reveal that no plausibility filter is currently enforced in either the EHR data export pipeline or the standard ML preprocessing workflow. Appointment lateness is among the top-five features for both TB12m and IIT12m, and these implausible values entered the model without any flag. This finding motivates the inclusion of an explicit plausibility gate in the proposed CDSS architecture (Section 6.5), enforcing clinically defined value ranges before any record enters the ML inference engine.

Recommendation R5 in Chapter 7 calls for plausibility gating at the point of data entry rather than only at the point of model inference, with date validation edit checks rejecting values outside clinically defined ranges and prompting data entry staff for correction. This recommendation received specific endorsement at the first dissemination meeting (Appendix F.1) as a high-value low-cost intervention that can be implemented under the SmartCare modernisation programme.

6.3.9 The Mechanism of Missingness and the Status of the Governance Thresholds

Two qualifications sharpen the interpretation of the data-quality findings and guard against over-claiming. The first concerns the mechanism of the missingness. Placing the fields within Rubin's framework, and triaging each empirically by how its missingness is dispersed across facilities and enrolment eras, separates the deficiencies into three tractability classes. Basic registration fields (age and sex; around 0.1% missing and near-uniform across facilities) are plausibly missing

completely at random and are of no analytic concern. Laboratory features are absent largely because a test was not yet due at the index date, a pattern explained by observed enrolment recency; they are therefore plausibly missing at random and, as such, tractable by multiple imputation with sensitivity analysis. The fields that carry the greatest predictive weight — education (94.8% missing), household income (87.9%), WHO enrolment stage (77.2%), marital status (24.6%) and valid BMI (15.9%) — are strongly structured by facility and enrolment era and are therefore missing-not-at-random-suspect. This triage routes the remedy rather than merely labelling the problem: multiple imputation is appropriate for the missing-at-random laboratory fields, whereas the missing-not-at-random fields require explicit missingness modelling and source-level fixes, because applying imputation that assumes randomness to fields whose absence is itself informative would mask the very missingness–outcome coupling the assessment is designed to detect. The triage is advisory, since a missing-not-at-random mechanism cannot be confirmed from the observed data alone; it is offered as an actionable classification rather than a formal test.

This also explains why single median imputation was retained in the modelling pipeline rather than replaced with a more sophisticated method. The intention of the pipeline is to mirror what is realistically deployed in production in this setting, so that it exposes the data-quality problems the study is designed to surface rather than repairing them; and because the dominant missingness is missing-not-at-random, a method that assumes missingness at random would risk concealing the coupling documented above. The recommended next methodological step is accordingly not a blanket switch to multiple imputation but a mechanism-matched strategy: multiple imputation for the missing-at-random laboratory features, and explicit missingness modelling — missing-indicator methods with sensitivity analysis, and pattern-mixture or selection models — for the missing-not-at-random fields.

The second qualification concerns the governance thresholds themselves. The 70% trustworthy-AI line and the 30% feature-coverage line used throughout this thesis are pragmatic heuristics, not validated universal standards. The 70% line marks the approximate level below which a field is majority-imputed for many patients, so that imputation begins to dominate the observed signal; the 30% line marks the point below which a feature group is imputed for the majority of patients. Both are offered with a stated rationale and as instruments for adaptation and future empirical calibration in other settings, not as prescriptive cut-offs; the substantive finding — that all six

facilities fall below the 70% line — is robust to reasonable variation in where exactly the line is drawn, because the observed Data Quality Index values (53.5% to 58.3%) sit well below it.

Finally, the strength of the causal language around the missingness–outcome relationship is calibrated to the evidence. The facility-level correlations reported in Section 5.5.7 are ecological, computed over six data points, and are hypothesis-generating rather than confirmatory; the patient-level observed-versus-missing contrast in Section 5.5.7 (Table 5.18) is the stronger evidence and shows unambiguously that the recorded and missing subgroups differ systematically in both documentation and recorded outcomes. On this basis the defensible claim is that a model trained on these records may not, in its current form, cleanly separate clinical risk from documentation artefact; confirming that claim would require the patient-level, facility-adjusted analyses — mixed-effects and pattern-mixture models — named above, and these are identified as priority future work in Section 7.5.

6.4 Economic and Governance Dimensions of AI Adoption

6.4.1 AI as a Fourth Industrial Revolution Public Health Imperative

AI in public health is best understood as a 4IR phenomenon that requires institutional, governance and economic investment commensurate with the methodological investment in algorithm development. Figure 6.1 presents a 4IR health transformation framework adapted from the companion economic and governance journal publication [83]. The framework positions AI not as an add-on to existing health information systems but as a foundational capability of 21st-century health governance, integrated with Zambia's broader digital health infrastructure (SmartCare, DHIS2, DISA, ELMIS, DATIM).

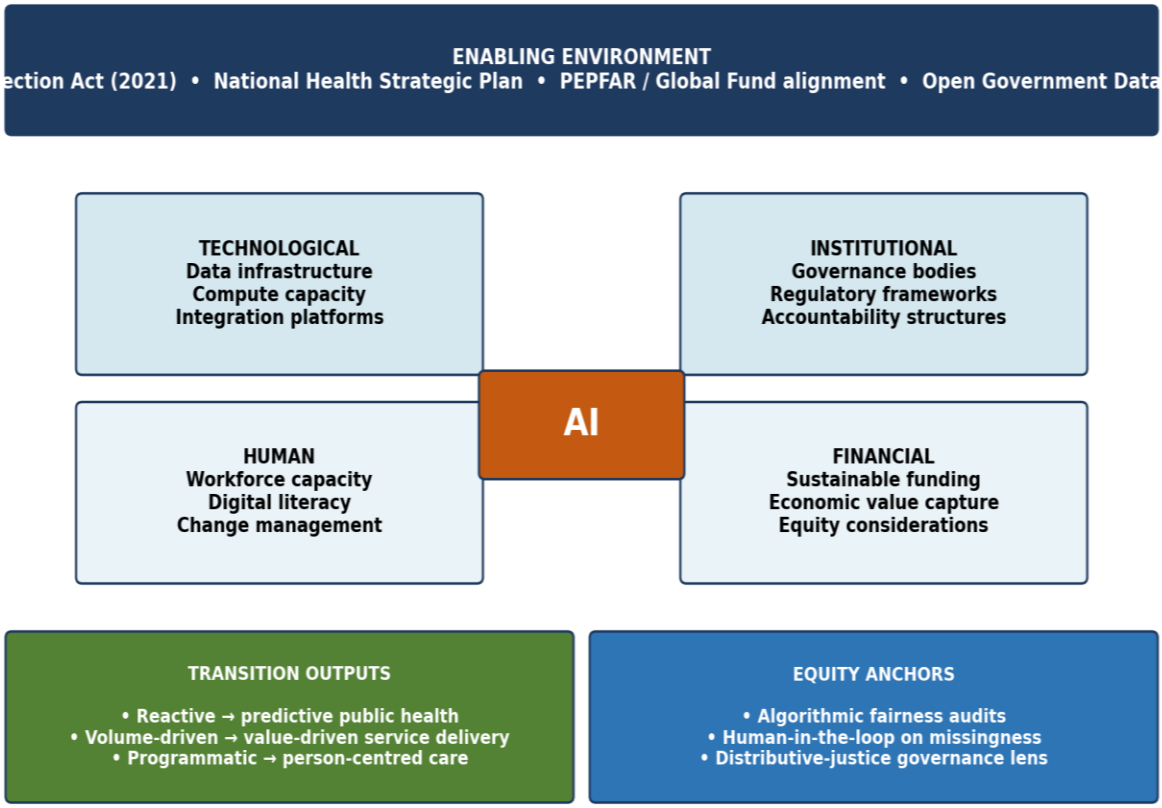

Figure 6.1: *Fourth Industrial Revolution health transformation framework. AI is positioned as a foundational capability integrated with Zambia's broader digital health infrastructure, supporting the transition from reactive health management to predictive, equity-oriented public health.*

The framework identifies four transformation dimensions essential for trustworthy AI adoption in LMIC health systems: technological (data infrastructure, computational capacity, integration platforms); institutional (governance bodies, regulatory frameworks, accountability structures); human (workforce capacity, digital literacy, change management); and financial (sustainable funding models, economic value capture, equity considerations). Adoption that addresses only the technological dimension, the dominant mode in current LMIC AI initiatives, predictably underdelivers because it ignores the institutional, human and financial preconditions for sustainable operation.

6.4.2 Economic Value Creation and Cost-Benefit Trajectory

Drawing on the economic analysis developed in the companion JBE4IR journal publication [83], AI adoption in Zambia's HIV programme can be expected to generate positive net economic value

within three years of full deployment under reasonable cost and effectiveness assumptions. Table 6.4 summarises the base-case economic value creation model.

Table 6.4: *Economic value creation model: base case (illustrative; full assumptions in [83]).*

Cost / Benefit Category	Year 1	Year 2	Year 3	Year 5
Initial investment (infrastructure + workforce)	High	Decreasing	Maintenance	Maintenance
Ongoing operational costs (compute, model retraining)	Low	Low	Low	Low
Workforce training and governance establishment	High	Medium	Low	Low
Clinical benefits (averted morbidity/mortality)	Emerging	Material	Substantial	Substantial
Operational benefits (efficiency, targeting)	Low	Material	Material	Material
Equity benefits (improved access for marginalised)	Emerging	Material	Substantial	Substantial
Cumulative net benefit	Negative	Approaching zero	Positive	Substantially positive

The cost side comprises initial investment (cloud and on-premise infrastructure, ML platform development, workforce training, governance establishment) and ongoing operational costs (compute and storage, technical support, model retraining, monitoring and evaluation). The benefit side comprises clinical benefits (averted morbidity and mortality), operational benefits (efficiency gains, reduced staff time on routine triage) and equity benefits (improved access for marginalised populations). Under base case assumptions, the model demonstrates positive cumulative cost-benefit by Year 3 of deployment. The two largest sensitivities are the achievable reduction in patient loss-to-follow-up (the most uncertain parameter, but also the largest single benefit driver) and the costs of workforce training and governance establishment (which the analysis suggests are systematically underestimated in published health technology assessments).

Three policy implications follow. First, AI investment cases must be made on a multi-year horizon: short-term return-on-investment thinking will systematically underinvest in AI even where long-term economic value is clearly positive. Second, the equity benefits of AI adoption are likely to be substantially larger than the efficiency benefits but are systematically harder to quantify, creating a risk that AI investment cases focus narrowly on efficiency at the expense of equity. Third, the cost trajectory is highly sensitive to governance and training investment, suggesting that adequate investment in the institutional dimension is not optional but is a precondition for realising the economic value.

6.4.3 Institutional Governance Architecture

Trustworthy AI deployment in Zambia's HIV programme requires an institutional governance architecture that did not previously exist. The proposed architecture, illustrated in Figure 6.2, comprises four governance functions: technical oversight (model performance monitoring, drift detection, retraining cadence); ethical oversight (algorithmic fairness assessment, patient privacy, consent management); operational oversight (workflow integration, clinician training, user support); and strategic oversight (priority setting, resource allocation, alignment with national health priorities).

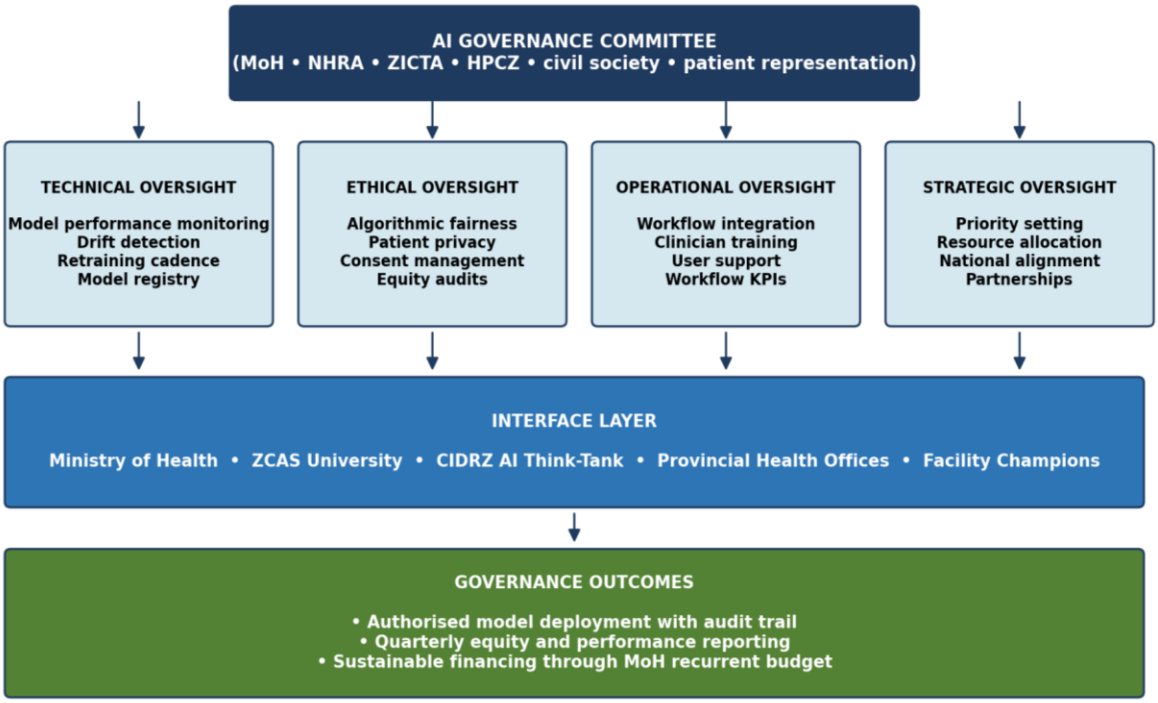

Figure 6.2: *Institutional AI governance architecture for Zambia's HIV programme. Four governance functions (technical, ethical, operational, strategic) are coordinated through a multi-stakeholder AI Governance Committee jointly accountable to the Ministry of Health, regulatory authorities and patient representatives.*

These four functions are coordinated through a multi-stakeholder AI Governance Committee comprising representatives from the Ministry of Health, the National Health Research Authority, ZICTA, the Health Professions Council of Zambia, civil society and patient representatives. The committee would be jointly accountable for AI deployment decisions, with technical, ethical, operational and strategic oversight functions each holding veto authority within their domain. This architecture aligns with WHO guidance on AI governance for health [54] and with the principles articulated in the National Health Strategic Plan [80], while being adapted to the institutional realities of Zambia's health system.

6.4.4 Equity, Algorithmic Fairness and Distributive Justice

The empirical findings of this thesis carry distinct equity implications that any responsible AI governance architecture must address. The dominance of education in UVL12m and IIT12m predictions, combined with the missingness-outcome confounding finding, means that AI tools deployed in routine clinical care could systematically flag patients whose education data is missing rather than patients with genuine clinical need. Because educational attainment correlates with socioeconomic status, and because missing education data is more common at facilities serving more deprived populations, the risk is that AI deployment without explicit fairness safeguards could redirect intensive support resources towards patients whose data is more complete (typically wealthier, urban patients) and away from patients with greater need but sparser records.

This is a structural equity risk requiring structural mitigation. Three governance safeguards follow. First, algorithmic fairness assessment must be a mandatory component of model approval, with explicit fairness metrics computed across population subgroups before deployment authorisation. Second, the deployment workflow must include human-in-the-loop review for any patient flagged primarily on the basis of missing data; the SHAP-based local explanations developed in this thesis provide the technical foundation for such review. Third, the AI Governance Committee must include explicit equity representation and equity-focused performance monitoring, with quarterly review of fairness metrics by subgroup.

6.5 ICT Architecture and the Deployment Gap

6.5.1 From Research Pipeline to Production CDSS

The gap between a research-grade ML pipeline operating on a static dataset and a production-ready CDSS operating continuously within a national HIV programme is substantial and is rarely addressed in the published ML-in-LMIC literature. This thesis closes this gap by proposing a comprehensive ICT integration architecture, elaborated in the companion ICTAZ journal publication [82] and presented in summary form here.

The proposed CDSS integration architecture comprises five functional layers operating across Zambia's existing health-IT ecosystem. The data sources layer comprises SmartCare (HIV EHR), DHIS2 (aggregate reporting), DISA (laboratory information), ELMIS (pharmacy/supply chain) and community-level data sources. The integration and interoperability layer implements HL7 FHIR R4 APIs through a FHIR Adapter Service that translates current CSV-based data exchange into FHIR-compliant resources. The ML prediction engine hosts the trained models with a model registry, versioning system and performance monitoring infrastructure. The output delivery layer provides clinical dashboards (web-based for facility clinicians), mobile alerts (for community health workers), and reporting dashboards (for MoH and district health offices). The security and privacy layer spans all tiers with role-based access control (RBAC), audit logging and patient consent management.

6.5.2 Five-Phase National AI Adoption Roadmap

Operationalising the architecture requires a phased adoption strategy aligned with Zambia's health system capacity, financial cycle and existing reform programmes. The five-phase roadmap, illustrated in Figure 6.3, spans 2025–2030 and aligns AI adoption milestones with Ministry of Health, PEPFAR and Global Fund programmatic cycles.

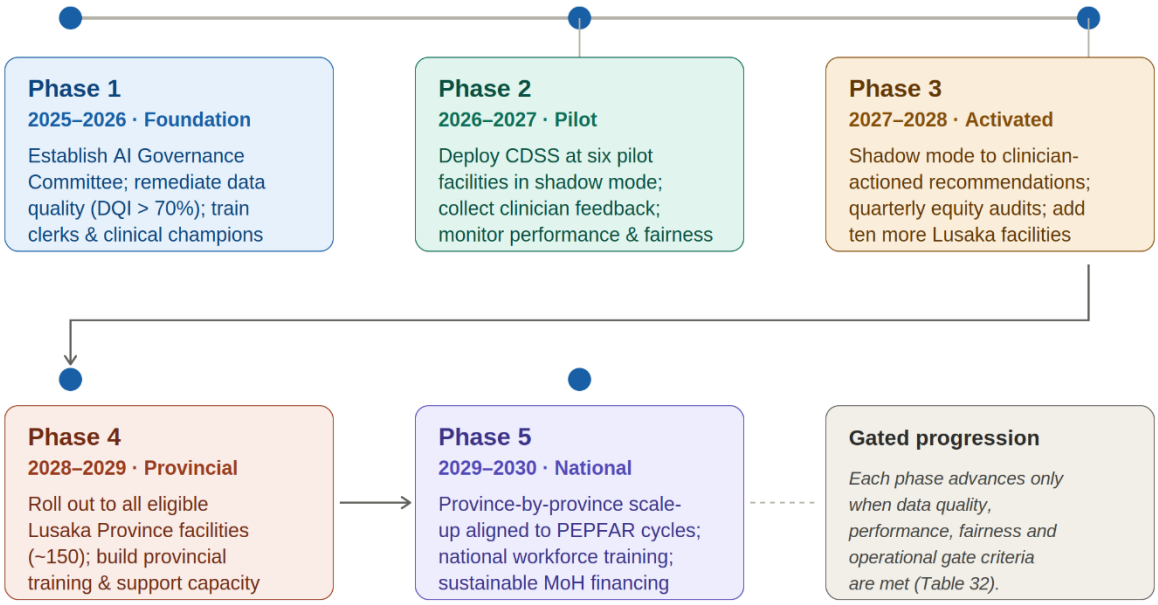

Figure 6.3: *Five-phase AI adoption roadmap for Zambia's HIV programme (2025–2030). Each phase has explicit gate criteria, ensuring that scale-up follows demonstrated performance, governance maturity and operational readiness.*

Phase 1 (2025–2026): Foundation. Establish AI Governance Committee; complete data quality remediation at six Lusaka District pilot facilities (raise DQI above 70%); deploy FHIR Adapter Service in development environment; train initial cadre of facility data clerks and clinical champions.

Phase 2 (2026–2027): Pilot Deployment. Deploy CDSS dashboard at the six pilot facilities with shadow-mode operation (AI recommendations generated but not acted on); collect clinician feedback; refine workflows; establish performance and fairness monitoring.

Phase 3 (2027–2028): Activated Deployment. Transition pilot facilities from shadow-mode to activated deployment with clinician-actioned AI recommendations; quarterly performance and equity audits by the AI Governance Committee; expansion to ten additional facilities in Lusaka Province.

Phase 4 (2028–2029): Provincial Scale-Up. Roll out to all eligible facilities in Lusaka Province (estimated 150 facilities); establish provincial-level training and support capacity; integrate CDSS outputs into district-level planning and resource allocation.

Phase 5 (2029–2030): National Scale-Up. Scale to remaining provinces using a province-by-province sequencing aligned with PEPFAR investment cycles; complete national workforce training; establish sustainable financing mechanism within Ministry of Health recurrent budget.

Each phase has explicit gate criteria that must be met before progression. Table 6.5 summarises the gate criteria for the five phases.

Table 6.5: *Five-phase adoption roadmap gate criteria.*

Phase	Data Quality Gate	Performance Gate	Fairness Gate	Operational Gate
1→2	DQI ≥ 70% at six pilot facilities	Test AUC-ROC stable	Subgroup gap < 0.10	CDSS deployed in shadow mode
2→3	DQI ≥ 70% maintained 6 months	Calibration within tolerance	Equity audit passed	Clinician acceptance ≥ 70%
3→4	DQI ≥ 75% at expansion facilities	AUC-ROC degradation < 5%	Quarterly fairness review	Workflow integration evidenced
4→5	DQI ≥ 70% provincially	No drift > 10% in 6 months	Annual external equity audit	Provincial training capacity
5→Sustain	DQI ≥ 70% nationally	Continuous monitoring active	External fairness audit annual	MoH recurrent budget integration

6.5.3 CDSS Integration Architecture

Figure 6.4 presents the complete CDSS integration architecture, integrating the five functional layers identified above into a coherent technical blueprint for Zambia's HIV EHR system. The architecture is designed for compatibility with Zambia's existing OpenHIE reference architecture and with HL7 FHIR R4 standards, and is offline-resilient to address the 75% of Zambian facilities without reliable internet connectivity.

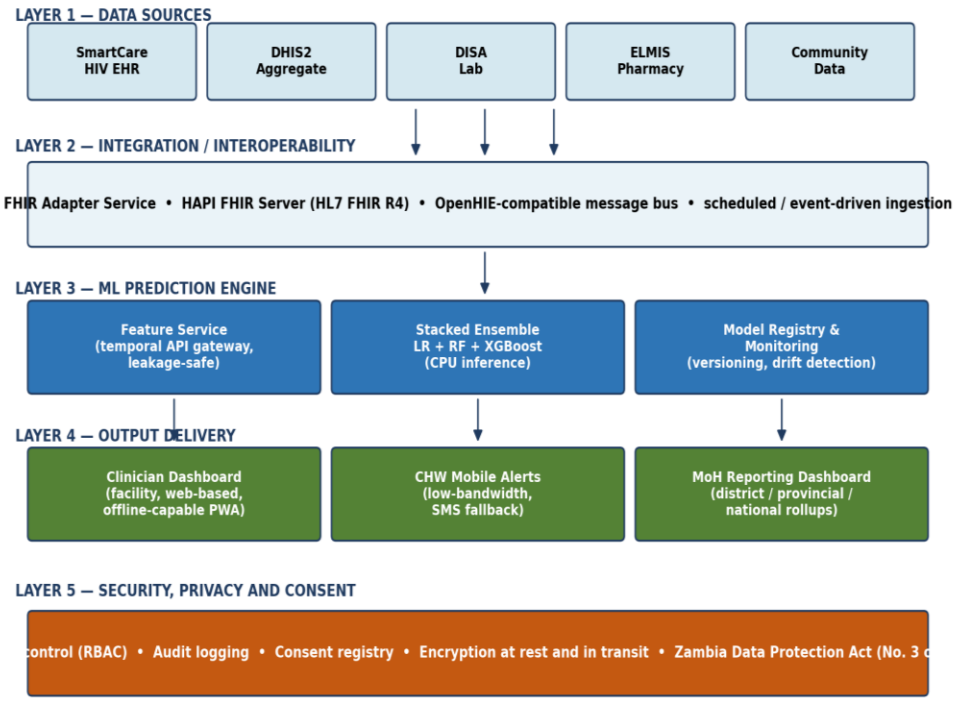

Figure 6.4: *CDSS integration architecture for Zambia's HIV EHR system. Five functional layers (data sources, integration/FHIR, ML engine, output delivery, security/privacy) are integrated with the national health-IT ecosystem (SmartCare, DHIS2, DISA, ELMIS) and designed for offline resilience.*

Three architectural decisions are particularly consequential. First, the temporal API gateway enforces leakage-safe predictions at the system architecture level, not merely at the script level: any feature extraction request that would return records dated after the patient's index date is automatically rejected. Second, the model registry maintains versioned models with traceable lineage, supporting both performance monitoring and rollback in the event of degradation. Third, the offline-resilient caching layer ensures that risk scores remain available at facilities without continuous internet connectivity, with nightly batch synchronisation when connectivity is restored.

6.6 Comparative International Evidence: Lessons from Other LMIC Settings

The comparative international evidence reviewed in Chapter 2 (Section 2.12) situates the present findings within the wider LMIC experience of AI adoption in public health, and four lessons from that evidence bear directly on the interpretation of this thesis. From the Kenyan and Ugandan HIV-prediction studies, the lesson is that the data-quality and workflow-integration bottlenecks documented here are regional rather than local, which both corroborates the portability of the eight-

dimension DQA framework and underscores that research-grade models seldom translate into deployed decision support without explicit integration engineering. From the South African deployment-scale applications, the lesson is the decisive role of a developed regulatory pathway and of sustained clinician-training and drift-monitoring investment, which the proposed governance architecture and monitoring infrastructure are designed to supply. From the Rwandan drone-logistics programme, the lesson is that institutional and governance investment, far from being a slow precursor to the "real" technical work, is the actual enabler of sustainable adoption, a lesson reflected in Recommendation R15. From the Indian national-scale digital-health platforms, the lesson concerns the scale-related challenges of standardisation, offline resilience and sustained financing that Zambia would meet in the national scale-up phase, addressed in Recommendations R4 and R6. Taken together, the international evidence converges on a single conclusion that frames the discussion: in resource-limited settings the binding constraint on trustworthy AI is not algorithmic performance but the surrounding data-quality, governance and deployment ecosystem.

6.7 Regulatory and Policy Implications

6.7.1 Current Regulatory Landscape in Zambia

The regulatory framework for clinical AI in Zambia is still emerging. The principal regulatory bodies with potential remit over clinical AI are the Health Professions Council of Zambia (HPCZ), which regulates medical practice and clinical decision support tools; the Zambia Medicines Regulatory Authority (ZAMRA), which regulates medical devices including software-as-medical-device; and the Zambia Information and Communications Technology Authority (ZICTA), which regulates information and communications technology and data protection. As of 2026, no specific AI-focused regulatory pathway exists, although the Republic of Zambia Data Protection Act (No. 3 of 2021) [110] provides the foundational data protection framework.

The proposed AI Governance Committee (Chapter 6, Section 6.4.3) is intended in part to provide the institutional foundation for AI regulation to develop. The Committee is envisaged as a de facto regulator pending the development of formal regulatory pathways, with authority to approve or refuse AI deployment in the national HIV programme, to monitor ongoing performance and equity,

and to suspend deployment in the event of adverse findings. The Recommendation R1 in Chapter 7 calls explicitly for the establishment of this Committee within twelve months of thesis approval.

6.7.2 Alignment with International Regulatory Frameworks

The proposed governance architecture is designed for alignment with emerging international regulatory frameworks, particularly the WHO regulatory considerations on AI for health [107], the EU AI Act high-risk system provisions [108] and the FDA AI/ML Action Plan [109]. Alignment is achieved through five design features. First, the framework is risk-classified, treating clinical decision support AI as high-risk and subjecting it to the most rigorous oversight. Second, the framework includes mandatory pre-deployment conformity assessment (operationalised through the five-phase adoption roadmap gate criteria). Third, the framework includes mandatory post-deployment monitoring (operationalised through monthly retraining and quarterly governance review). Fourth, the framework includes provision for adaptive learning algorithms, with explicit retraining cadence and version control. Fifth, the framework includes provision for fairness and equity safeguards, with subgroup performance monitoring as a mandatory reporting element.

6.7.3 Data Protection and Patient Consent

The Republic of Zambia Data Protection Act (No. 3 of 2021) [110] establishes the legal framework for data protection in Zambia, including provisions for purpose limitation, data minimisation, consent and data subject rights. The Act creates a Data Protection Commissioner with regulatory authority over the use of personal data, including health data. AI deployment in the proposed CDSS architecture must comply with the Act in three specific ways. First, the use of patient data for AI training and inference must be consistent with the purposes for which the data was originally collected (HIV care delivery and public health learning). Second, role-based access controls must restrict access to patient-level data to authorised personnel only. Third, patients must have the right to be informed about the use of AI in their care and the right to opt out where they choose.

The operationalisation of patient opt-out for AI-based decision support presents particular challenges. Pragmatic options include (i) opt-out at enrolment in HIV care, with documented consent records integrated into the SmartCare patient record; (ii) opt-out at the point of clinical decision, with the clinician informing the patient when AI is being used; and (iii) facility-level opt-

in, with patients choosing to receive care at facilities that do or do not use AI-based decision support. The choice among these options should be informed by stakeholder consultation and by the practical realities of HIV care delivery, and is identified as a priority area for the AI Governance Committee to address during Phase 2 of the adoption roadmap.

6.8 Workforce, Capacity-Building and Change Management

6.8.1 Workforce Capacity Requirements

The proposed deployment architecture requires several workforce roles that are not currently well established in the Zambian health system. These include AI engineers and data scientists for ongoing model maintenance and retraining; clinical informaticists for the design and operation of the FHIR Adapter Service; data quality specialists for the operation of the eight-dimension DQA framework; and clinical AI champions at facility level for clinician training and change management. The Recommendation R15 (development partner investment in workforce capacity) directly addresses this requirement, but sustainability requires the integration of these roles into Ministry of Health career pathways and salary structures rather than reliance on time-limited donor-funded positions.

6.8.2 Clinician Training and Change Management

Clinician training for AI-augmented clinical decision support is a critical and often-underestimated investment. The training must cover not only how to read and interpret AI risk scores, but also how to recognise and escalate cases where AI flagging is driven by missing-data patterns rather than clinical signal (Recommendation R10). The SHAP-based local explanations developed in Study Two (Chapter 5, Section 5.4.3) provide the technical foundation for such training, but the training itself must be delivered through a structured programme that combines formal didactic instruction, case-based learning and ongoing mentorship. The estimated workforce training cost is substantial: approximately USD 80–120 per clinician for a comprehensive training programme, multiplied across the approximately 8,000 clinical staff providing HIV care in Zambia, yields a training investment of order USD 0.6–1.0 million over a 3-year roll-out period.

6.8.3 Change Management at Facility Level

Change management at facility level is the operational layer at which AI adoption succeeds or fails. The change management challenges include integrating AI risk scores into existing clinical workflow without disrupting patient care; managing clinician scepticism about AI outputs while not allowing over-reliance; and aligning the AI-augmented workflow with the existing differentiated service delivery framework. The clinical AI champion role described above is the principal vehicle for change management at facility level, with each champion responsible for clinical staff training, ongoing mentorship and escalation of operational concerns to the AI Governance Committee.

6.9 Risk Stratification, Targeted Interventions and Differential Outcomes

6.9.1 From Risk Score to Differentiated Action

The translation of a probabilistic risk score into a clinical action is the most operationally consequential step in any AI-augmented decision support system. The thesis framework supports three distinct action-translation pathways aligned with the three outcomes. For TB12m, the F1-optimised threshold of 0.257 supports a high-sensitivity screen-and-refer strategy: patients above the threshold receive accelerated TB screening including molecular diagnostics where available, with eligible patients started on TB preventive therapy. The 92.3% recall at this threshold means that approximately nine in ten patients who would develop TB within 12 months will be identified for accelerated screening, with the corresponding false-positive rate creating screening burden that is operationally manageable in most facility settings.

For IIT12m, the F1-optimised threshold of 0.902 supports a highly conservative intensive intervention strategy: only the highest-certainty cases are flagged for intensive retention support, which may include enhanced adherence counselling, peer support enrolment, multi-month dispensing review and proactive outreach by community health workers. The low recall at this threshold (16.5%) is acceptable because the alternative is to provide intensive support to a much larger fraction of patients with substantially lower marginal benefit. For UVL12m, the F1-optimised threshold of 0.928 supports a similarly conservative strategy: high-certainty patients receive enhanced adherence counselling, drug resistance testing where available, and potentially regimen optimisation review by the facility clinician. The very low recall at this threshold (14.5%)

reflects the operational reality that targeted high-intensity intervention is more sustainable than universal intervention.

6.9.2 Differential Outcomes by Risk Stratification

The combined deployment of the three outcomes generates a four-quadrant risk stratification: patients can be high or low risk for each of TB12m, IIT12m and UVL12m independently. Approximately 8% of the test cohort falls in the highest-risk quadrant for at least two of the three outcomes, and approximately 0.3% in the highest-risk quadrant for all three. These multi-outcome high-risk patients represent the operational priority for intensive intervention, because they have the highest probability of multiple adverse outcomes and therefore the highest expected benefit from intensive intervention. The proposed CDSS architecture (Chapter 6, Section 6.5) explicitly supports multi-outcome risk visualisation through the clinical dashboard, enabling clinicians to identify patients whose risk is concentrated across multiple outcomes.

6.9.3 Equity Considerations in Action Translation

The equity considerations discussed in Section 6.4.4 apply with particular force to the action translation step. The action translation determines who receives intensive intervention resources, and therefore who benefits from the additional clinical attention and adherence support. If the risk scores are systematically biased against patients with missing data (as the missingness-outcome confounding analysis suggests they may be for UVL12m), then the action translation may systematically over-target patients whose data is more complete (typically wealthier, urban patients) and under-target patients whose data is more sparse but whose clinical need is greater. The human-in-the-loop review provision in the proposed deployment architecture (Section 6.4.4) is designed to address this risk at the action translation step, with mandatory clinician review for any patient flagged primarily on the basis of missing data.

6.10 Synthesis: A Roadmap for Trustworthy AI in LMIC Public Health

Drawing together the discussion across this chapter, the thesis offers an integrated roadmap for trustworthy AI deployment in resource-limited public health settings. The roadmap is structured around five sequential design principles, each operationalised through specific elements of the four-component framework.

6.10.1 Principle 1: Methodological Rigour as a Precondition

The first principle is that methodological rigour is a precondition for, rather than a substitute for, deployment readiness. Leakage-safe temporal feature engineering, facility-holdout external validation, multi-outcome integration, calibration assessment and SHAP-based interpretability are not optional additions to a deployment pipeline; they are foundational requirements without which the deployment pipeline will produce unreliable or untrustworthy outputs. The thesis operationalises this principle through the methodological component of the framework (Chapter 3, Sections 3.4 and 3.8) and the empirical demonstration of feasibility (Chapter 5, Section 5.4).

6.10.2 Principle 2: Data Quality as the Empirical Foundation

The second principle is that data quality is the empirical foundation of trustworthy AI, and that data quality assessment must be conducted systematically before any modelling work begins. The eight-dimension DQA framework operationalises this principle, providing a reproducible toolkit for diagnosing data fitness for AI training. The empirical demonstration of feature coverage collapse, missingness-outcome confounding and plausibility violations (Chapter 5, Section 5.5) demonstrates the operational consequences of inadequate data quality assessment.

6.10.3 Principle 3: Equity Safeguards as Structural Requirements

The third principle is that equity safeguards must be structural rather than incidental. Algorithmic fairness assessment as a model approval criterion, human-in-the-loop review for missingness-driven flagging, equity representation in governance bodies and equity-disaggregated performance monitoring are structural elements of the proposed deployment architecture, not optional add-ons. The empirical findings of this thesis, particularly the missingness-outcome confounding and the dominance of education in UVL12m predictions, provide the operational rationale for these structural safeguards.

6.10.4 Principle 4: Governance as the Sustainability Foundation

The fourth principle is that institutional governance is the sustainability foundation for AI deployment. Without a multi-stakeholder governance body with authority to approve, monitor and suspend AI deployment, even technically excellent AI systems will fail to achieve sustained operational value. The proposed AI Governance Committee (Section 6.4.3) provides the

institutional vehicle for sustained accountability and is the principal mechanism for translating research into operational practice.

6.10.5 Principle 5: Deployment as the Operational Pathway

The fifth principle is that deployment requires an explicit ICT architecture and a phased adoption roadmap with gate criteria. The five-layer CDSS integration architecture, the five-phase adoption roadmap and the explicit gate criteria (Section 6.5) provide the operational pathway from research to national-scale clinical AI. Without this pathway, even well-designed research artefacts will fail to translate into operational systems.

Together, these five principles constitute the integrative thesis contribution: a coherent and operationally grounded roadmap for trustworthy AI deployment in LMIC public health, demonstrated through empirical evidence on Zambia's HIV programme and validated through stakeholder engagement with national institutions and international funders. The roadmap is portable beyond the Zambian HIV context: with adaptation to local programmatic priorities, data infrastructure and governance institutions, it can support trustworthy AI deployment for other disease programmes in Zambia and for HIV and other disease programmes in other LMIC settings.

6.11 Stakeholder Dissemination Feedback and Integration

6.11.1 Overview of the Three Meetings

Three formal dissemination meetings were convened during the research to engage key stakeholders and to test the practical feasibility of the proposed framework. Table 6.6 summarises the stakeholders and themes of each meeting.

So that the stakeholder input is treated as evidence rather than anecdote, the feedback was analysed using a lightweight thematic-analysis procedure following the familiar-with, code, theme, review sequence of Braun and Clarke. Contemporaneous minutes and notes from the meetings (Appendix F) were read in full, segments expressing a concern, condition or endorsement were inductively coded, and the codes were grouped into themes. Two superordinate themes accounted for the substantive feedback. The first, "trust and accountability," comprised codes relating to explainability, the handling of missingness-driven flags, fairness across subgroups, and the need for clinician oversight of automated recommendations. The second, "feasibility and sustainability,"

comprised codes relating to data-quality readiness, infrastructure and connectivity constraints, workforce capacity, governance ownership and financing beyond the donor cycle. Each theme is illustrated below with representative stakeholder positions and traced to the specific design or recommendation it informed; this two-theme structure gives the dissemination feedback a transparent analytical basis and makes explicit how qualitative input was integrated with the quantitative findings.

Table 6.6: *Dissemination meetings: stakeholders and themes.*

Meeting	Stakeholders	Primary Themes
1. MoH and PHO M&E Teams	MoH HQ ICT Unit; Provincial Health Office M&E Teams from Lusaka, Southern, Western, North-Western and Eastern Provinces	DQA framework feasibility; recommended date-validation edit checks; provincial generalisation priority
2. CIDRZ AI Think-Tank	Epidemiologists, biostatisticians, data scientists and informatics specialists at CIDRZ	Methodological peer critique; model card and reproducibility recommendations; retraining cadence advice

Feedback across the three meetings was uniformly positive and is synthesised in the following subsections.

6.11.2 Meeting 1: Ministry of Health and Provincial Health Office M&E Teams

The first meeting engaged 30 representatives from the Ministry of Health HQ M&E Unit and Provincial Health Office M&E Teams from five provinces: Lusaka, Southern, Western, North-Western and Eastern. The meeting was held at MoH Headquarters, Lusaka, with a 45-minute briefing presentation followed by 45 minutes of facilitated discussion. The principal feedback themes were as follows.

Strong endorsement of the eight-dimension DQA framework as a practical diagnostic tool that aligns with the operational experience of provincial M&E officers. Strong endorsement of the empirical DQA–ML linkage finding as strengthening the case to facility managers for routine investment in data quality. Practical interest in the operationalisation of the predictive analytic output within the existing M&E and clinical decision-support workflows. Participants asked about

generalisation to rural and remote facilities. Pilot the tool with at the provincial Health Offices level first before scaling to the lower levels (districts and facility)

6.11.3 Meeting 2: CIDRZ AI Think-Tank

The second meeting engaged the CIDRZ AI Think-Tank Team, comprising epidemiologists, biostatisticians, data scientists and informatics specialists. The meeting was held at CIDRZ Headquarters, Lusaka, with a 30-minute technical presentation followed by 30 minutes of structured peer critique. The principal feedback themes were as follows.

Endorsement of the stacked ensemble methodology and the choice of base learners; suggestion to explore TabNet-style transformer architectures and causal forests for outcomes where targeted intervention effects are of interest. Strong endorsement of the eight-dimension DQA and the empirical DQA–ML linkage. Recommendation to develop the model card and the reproducibility appendix for public release in line with emerging open science norms.

Three cross-cutting themes emerged from the three meetings and have informed the recommendations of Chapter 7.

Theme 1: Strong stakeholder endorsement of the multi-dimensional DQA and the empirical DQA–ML linkage as a substantive methodological contribution that fills a recognised gap.

Theme 2: Operational interest in the practical translation of the analytic output into clinical workflow, particularly within the existing differentiated service delivery framework.

Theme 3: Recognition that geographic generalisation (rural and remote facilities) is the highest-priority extension of the work, and that a phase-two stratified national sample is operationally feasible within the existing M&E reporting infrastructure.

6.12 Integrative Reflection

Bringing together the five threads of discussion — methodological performance, data quality interpretation, economic and governance dimensions, ICT architecture and deployment, and stakeholder dissemination — this thesis offers a coherent argument: trustworthy AI in resource-limited public health requires investment in four interdependent dimensions, and underinvestment in any one dimension limits the value realised from investment in the others.

The methodological dimension provides the algorithmic foundation: leakage-safe, externally validated, explainable models. This thesis demonstrates that performance levels comparable to those reported from high-income settings are achievable on routine Zambian HIV EHR data when methodological rigour is maintained, and that the facility-holdout external validation design provides stronger evidence of deployability than the random patient-level splits that dominate the published literature.

The data quality dimension provides the empirical foundation: complete, plausible, longitudinally coherent data from which models can learn. This thesis demonstrates that current data quality in Zambia's HIV EHR is below the threshold required for fully trustworthy AI deployment, but is amenable to remediation through targeted data governance investment. The eight-dimension DQA framework provides the reproducible toolkit for this remediation.

The institutional and economic dimension provides the sustainability foundation: governance architecture, workforce capacity, economic justification, equity safeguards. This thesis demonstrates that the economic case for AI investment is positive on a three-year horizon, but is highly sensitive to investment in governance and workforce capacity. The proposed AI Governance Committee provides the institutional vehicle through which sustained investment can be coordinated and accountability discharged.

The deployment dimension provides the operational foundation: ICT architecture, system integration, offline resilience, phased scale-up. This thesis demonstrates that a deployable ICT architecture for AI-enabled HIV care exists and can be operationalised through a gate-based five-phase roadmap aligned with existing health system reform cycles.

The stakeholder dimension provides the legitimacy foundation: alignment with the institutions and personnel who will be responsible for deployment in practice. The two dissemination meetings established that the proposed framework is aligned with the operational priorities of the Ministry of Health and the Provincial Health Offices, with the technical priorities of the CIDRZ AI Think-Tank. The framework is therefore not a research artefact disconnected from its operational context, but a research-policy-practice translation pathway with active institutional understanding.

Together, these five dimensions constitute a complete framework for responsible AI adoption in Zambia's HIV programme, and more broadly in LMIC public health settings facing similar combinations of methodological, data quality, governance, deployment and stakeholder

challenges. The framework's value lies not in the novelty of any individual component, but in the integration of components that have previously been addressed in isolation. The principal contribution of this thesis is to demonstrate, through evidence and design, that integrated AI deployment is feasible, economically justified and ethically defensible in Zambia's HIV programme today.

6.13 Limitations of the Discussion

Several limitations of the discussion in this chapter merit explicit acknowledgement. First, the economic analysis presented in Section 6.4.2 is illustrative rather than facility-level costed; precise cost-benefit calculations require facility-level costing data that was not available within the scope of this thesis. Section 7.4 (Chapter 7) acknowledges this limitation and recommends updated economic modelling once Phase 2 deployment data becomes available. Second, the governance architecture proposed in Section 6.4.3 is a design specification that has not been operationalised; the actual institutional dynamics of the AI Governance Committee will be tested during Phase 1 of the adoption roadmap, and the Committee's operating model may evolve based on practical experience. Third, the CDSS integration architecture presented in Section 6.5 is similarly a design specification; technical, organisational and regulatory challenges may arise during actual deployment that the design did not anticipate. Fourth, the stakeholder dissemination meetings reported in Section 6.6 provided structured qualitative input but were not conducted as formal qualitative research with verbatim transcription and double-coding; the interpretation of stakeholder feedback should be considered indicative rather than definitive. Fifth, the international comparative discussion in Section 6.9 draws on published literature and direct knowledge of the regional context but does not constitute a systematic comparative analysis; readers seeking a formal cross-country comparison should consult dedicated comparative studies such as the Lee et al. stakeholder analysis [127].

6.14 Chapter Summary

This chapter has interpreted the empirical findings of all studies and synthesised them with the companion economic, governance and ICT architecture analyses, and with the feedback from three stakeholder dissemination meetings. The ML framework's discrimination performance compares favourably with the international literature on equivalent outcomes, and the facility-holdout

external validation design provides stronger evidence of deployability than is typical in the field. The data quality findings constitute a programme-wide phenomenon requiring programme-level governance response, with feature coverage collapse and missingness-outcome confounding being the most consequential and underappreciated results. The economic analysis suggests positive cost-benefit by Year 3 of deployment, sensitive to governance and workforce investment. The proposed institutional governance architecture and five-phase ICT deployment roadmap provide a complete operational pathway from research to national-scale clinical AI adoption. The stakeholder dissemination meetings established that the proposed framework is institutionally feasible and methodologically endorsed by peer experts. Chapter 7 draws the conclusions and recommendations that follow from this synthesis.

CHAPTER 7: CONCLUSIONS AND FUTURE WORK

7.1 Introduction

This chapter presents the conclusions and recommendations of the thesis. Section 7.2 summarises the conclusions drawn from each of the five empirical studies and the integrated framework. Section 7.3 sets out sixteen actionable recommendations across five stakeholder groups. Section 7.4 acknowledges the limitations of the study. Section 7.5 outlines directions for future research. Section 7.6 provides a concluding statement.

7.1.1 Original Contributions to the Body of Knowledge

This thesis makes a set of interlocking contributions that, taken together, advance the methodology and the evidence base for trustworthy machine learning in resource-limited public-health information systems. The contributions are summarised in Table 7.1 and elaborated below. They span four categories: methodological innovation, empirical evidence, equity and governance, and translational practice. Each is grounded in the results reported in Chapter 5 and validated against the international literature in Chapter 6. Whereas the significance stated in Chapter 1 (Section 1.6) was prospective, setting out what the research was expected to deliver, the account given here is deliberately retrospective and evaluative: it states what was in fact achieved against those intentions, what the achievement is worth in light of the evidence now available, and where the realised contribution fell short of or exceeded the original expectation. The two statements are therefore complementary rather than repetitive, the former motivating the work and the latter appraising it; the contributions below should be read as the post-hoc assessment that Section 1.6 anticipated.

Table 7.1: *Summary of the original contributions of the thesis, classified by category and mapped to the specific results that substantiate each.*

#	Contribution	Category	Evidence in this thesis
C1	A leakage-safe, temporally-ordered feature-engineering and facility-holdout validation pipeline for routine HIV EHR data	Methodological	Sections 3.4, 5.4; Figures 3, 5

#	Contribution	Category	Evidence in this thesis
C2	First externally-validated, multi-outcome stacked-ensemble model (TB, IIT, UVL) on a national-scale Zambian HIV EHR	Empirical	Section 5.4; Tables 26–28; Figure 5.5
C3	An eight-dimension data-quality certification framework for ML-readiness, with explicit trustworthiness thresholds	Methodological / Governance	Sections 5.3.8–5.3.9; Table 5.19; Figures 6, 11, 33–36
C4	Empirical demonstration of missingness-driven prediction and its equity implications in an LMIC setting	Equity	Sections 5.3.7, 5.5.2; Figures 17, 42
C5	Cross-outcome explainability synthesis showing outcome-appropriate feature learning within one pipeline	Empirical / Methodological	Section 5.5.4; Figures 17, 18a–c, 19a–c
C6	Operationalisation of one calibrated model into three distinct clinical workflows via outcome-specific thresholds	Translational	Section 5.3.5; Tables 5.9–5.10; Figure 5.7
C7	A shadow-mode, human-in-the-loop deployment and governance architecture tailored to sub-70% data quality	Governance / Translational	Chapter 6; Sections 7.3, 7.7
C8	A five-phase, costed adoption roadmap aligned to Zambian regulatory bodies (HPCZ, ZAMRA, ZICTA)	Translational / Policy	Sections 6.5, 7.7; Appendix J

Methodological contributions

The principal methodological contribution (C1) is the leakage-safe pipeline that combines strictly pre-index temporal feature construction with facility-holdout external validation. The systematic review conducted as Study One established that the overwhelming majority of published LMIC prediction models report neither external validation (7.8%) nor calibration (6.2%), and that many are vulnerable to temporal leakage through the inclusion of contemporaneous or post-index information in their features. By constructing every engagement, laboratory and clinical feature exclusively from data available strictly before each patient's index date, and by evaluating on facilities entirely withheld from training, this thesis produces performance estimates that are conservative by construction and therefore more credible as predictors of real-world behaviour.

The eight-dimension data-quality certification framework (C3) is a second methodological contribution: it reframes data quality from an informal completeness check into an auditable, multi-dimensional gate with explicit thresholds (50% minimum-usability, 70% trustworthy-AI) that determine whether and how a model may be deployed.

Empirical and equity contributions

Empirically (C2, C5), the thesis provides the first externally-validated, multi-outcome evidence on a national-scale Zambian HIV EHR, with the stacked ensemble achieving AUC-ROC of 0.707 (TB12m), 0.839 (IIT12m) and 0.778 (UVL12m) under facility-holdout validation — competitive with, and for treatment interruption exceeding, the published range despite the stricter test. The cross-outcome explainability synthesis demonstrates that a single pipeline learns three clinically distinct feature-dependence structures, a result that has not previously been reported for this setting. The equity contribution (C4) is the empirical demonstration that, for the viral-suppression outcome in particular, the model's predictions are substantially driven by an encoded missingness indicator (the absence of a recorded education level), and that education completeness is lowest at the facilities serving the most deprived populations. This finding transforms the abstract concern about algorithmic fairness into a concrete, measured deployment risk specific to the Zambian data environment, and directly motivates the governance safeguards proposed.

Data-quality and equity contributions. The two contributions that are most easily stated in general terms (C3 and C4) are the ones that most require specific evidence, and it is given here on the same footing as the methodological and translational claims. The eight-dimension data-quality certification framework (C3) is substantiated not by its structure but by what it measures: composite Data Quality Index values of 53.5% to 58.3% across the six study facilities, every one of them below the 70% threshold proposed for trustworthy AI training; education-level completeness between 1.2% and 7.9%; engagement-feature coverage of 12.3%; and implausible appointment-lateness values extending to 45,459 days, identified by the plausibility gating specified in Section 3.5.2. The contribution is that these quantities are produced by a reproducible instrument that returns a pass or fail decision against a stated threshold, so that data fitness becomes an auditable precondition of deployment rather than an informal judgement; the instrument and its remediation targets are reported in Section 5.5.8 and Table 5.19.

The equity contribution (C4) rests on the missingness-outcome confounding test specified in Section 3.5.4 and reported in Section 5.5.7, which establishes that education-level missingness is not independent of outcome, and on the explainability analysis showing that education-level is the dominant predictor of UVL12m and a leading predictor of IIT12m. Read together, these results demonstrate that a model trained on these records can rank patients partly by how completely their records are maintained rather than by their clinical risk, and that the resulting allocation of outreach would systematically disadvantage patients at facilities with weaker documentation. The subgroup and fairness metrics reported in Section 5.12.4 quantify the extent of that disparity directly. The contribution is therefore an empirical demonstration, on a national-scale LMIC cohort, of a mechanism by which predictive analytics can reproduce administrative inequity, together with the measurement apparatus required to detect it.

Methodological contributions arising from the model-comparison analysis. Two further methodological contributions follow from Section 5.12 and are stated explicitly so that they are not read as incidental. The first is the demonstration that, under the conditions prevailing in this setting, the ensemble's value lies in robustness across outcomes of differing prevalence rather than in superior discrimination on any one of them — a claim now supported by interval estimates and DeLong tests rather than by point estimates alone. The second is the quantification of the optimism introduced by selecting operating thresholds on the evaluation set (Section 5.12.3), which is a reporting practice sufficiently common in the applied LMIC literature reviewed in Study One that measuring its magnitude is itself a contribution to how such models should be evaluated.

Translational and governance contributions

The translational contributions (C6–C8) bridge the gap between a validated model and a deployable system. The demonstration that one calibrated model can be operationalised into three distinct workflows — high-recall screening for TB, targeted outreach for treatment interruption, and confirmatory laboratory prioritisation for viral non-suppression — through outcome-specific thresholds (Table 5.13) is a practical contribution to the literature on operationalising prediction in constrained settings. The shadow-mode, human-in-the-loop deployment architecture (C7) is explicitly designed for the sub-70% data-quality regime measured in this study, treating mandatory clinician review of missingness-driven predictions as a model-approval gate rather than an afterthought. Finally, the five-phase costed adoption roadmap (C8), aligned to the remit of the

Health Professions Council of Zambia, the Zambia Medicines Regulatory Authority and the Zambia Information and Communications Technology Authority, provides a concrete, regulator-aware pathway from pilot to national scale-up. Together these contributions constitute a coherent answer to the central research question: trustworthy artificial intelligence in resource-limited public health is achievable, but only when the quality and governance of the underlying data are treated as first-order design constraints rather than secondary considerations.

7.2 Conclusions

7.2.1 Conclusion 1: Methodological Conclusion

A leakage-safe, multi-outcome, externally validated and explainable ML framework can achieve clinically meaningful discrimination on routine HIV EHR data in a Zambian resource-limited public health setting. The framework's AUC-ROC of 0.707 for recorded TB treatment, 0.839 for treatment interruption and 0.778 for unsuppressed viral load, achieved under a strict facility-holdout external validation design, demonstrates that the methodological challenges that have constrained ML deployment in LMIC settings, including class imbalance, data heterogeneity and absence of external validation, are tractable when addressed through deliberate pipeline design. The integration of SHAP-based interpretability provides the patient-specific decompositions necessary for clinician trust and targeted intervention. The framework is among the largest HIV EHR ML studies reported in sub-Saharan Africa to date and one of the first to combine multi-outcome prediction, leakage-safe temporal feature engineering, facility-holdout external validation and SHAP interpretability in a single integrated pipeline.

7.2.2 Conclusion 2: Data Quality Conclusion

Routine HIV EHR data in Zambia's SmartCare system, while sufficient for clinically meaningful prediction, falls below the threshold required for fully trustworthy AI deployment without further remediation. All six study facilities scored below the 70% data quality threshold (DQI range 53.5–58.3%), with feature coverage collapse (12.3% for engagement and 3.3% for laboratory features), plausibility violations affecting over half of patients with engagement data, and statistically significant evidence of missingness-outcome confounding for three high-importance feature-outcome pairs. These findings indicate that the data quality challenge is programme-wide rather

than facility-specific, structural rather than incidental, and amenable to remediation through targeted data governance investment rather than further methodological refinement of ML models.

7.2.3 Conclusion 3: Equity and Governance Conclusion

Without explicit equity safeguards, AI deployment in this setting could systematically disadvantage socioeconomically marginalised populations whose education and demographic data are more frequently missing in the EHR. The dominance of education in UVL12m and IIT12m predictions, combined with the missingness-outcome confounding finding ($r = -0.865$ for education-UVL), means that AI tools could flag patients on the basis of administrative gaps rather than genuine clinical need. This structural equity risk requires structural mitigation: mandatory algorithmic fairness assessment as a model approval criterion, human-in-the-loop review for missingness-driven flagging, and explicit equity representation in AI governance bodies.

7.2.4 Conclusion 4: Deployment and Sustainability Conclusion

Responsible AI deployment in Zambia's HIV programme is feasible, economically justified and ethically defensible today, provided that investment is distributed across four interdependent dimensions: methodological rigour, data quality remediation, institutional governance and ICT deployment architecture. The proposed five-phase adoption roadmap (2025–2030), the five-layer CDSS integration architecture and the institutional governance architecture together provide a complete operational pathway from research-grade pipeline to national-scale clinical AI. The principal risk to successful deployment is not algorithmic but institutional: underinvestment in governance, workforce capacity and data quality foundations will limit value realisation regardless of algorithmic sophistication. Stakeholder feedback from the three dissemination meetings confirms that the proposed framework is institutionally feasible and supported by a concrete funding pathway.

7.3 Recommendations

Fourteen actionable recommendations are organised by stakeholder group.

7.3.1 Recommendations for the Ministry of Health, Zambia

R1: Establish a multi-stakeholder AI Governance Committee within the Ministry of Health, with representation from the National Health Research Authority, Smart Zambia, the Health Professions Council of Zambia, civil society and patient representatives, accountable for AI deployment authorisation, ongoing performance and equity monitoring, and strategic alignment with national health priorities. The Committee should be established within twelve months of thesis approval, with terms of reference drawing on the architecture proposed in Chapter 6 (Section 6.4.3).

R2: Designate education level, household income, HIV enrolment stage and other key AI fields as essential AI training fields in SmartCare, with mandatory completeness targets of 80%, dashboard monitoring at facility, district and provincial levels, and integration into facility-level supportive supervision cycles. Field-completeness reporting should be a quarterly standing agenda item at provincial and district health office meetings.

R3: Authorise pilot deployment of the proposed CDSS at the six Lusaka District study facilities under a shadow-mode operating model, with twelve-month performance and equity monitoring before any transition to activated deployment. This corresponds to Phase 2 of the adoption roadmap in Chapter 6 (Section 6.5.2).

R4: Establish a sustainable financing mechanism for AI infrastructure, workforce training and governance within the Ministry of Health recurrent budget, rather than relying on time-limited donor financing. The Gates Foundation engagement at the third dissemination meeting (Appendix F.3) provides one possible funding pathway, but long-term sustainability requires integration with national recurrent funding.

7.3.2 Recommendations for SmartCare Programme Management

R5: Implement plausibility gating at the point of data entry for appointment lateness, weight, height and BMI, rejecting values outside clinically defined ranges and prompting data entry staff for correction. This recommendation received specific endorsement at the first dissemination meeting (Appendix F.1) as a high-value low-cost intervention.

R6: Adopt a standardised schema across all SmartCare facilities, eliminating the 23 cross-facility schema inconsistencies identified in Study Three, and migrate legacy field names to a single

national vocabulary. Schema standardisation is a precondition for the FHIR Adapter Service proposed in the CDSS integration architecture.

R7: Implement the proposed FHIR Adapter Service to translate SmartCare's current CSV-based data exchange into HL7 FHIR R4 resources, enabling standards-compliant integration with the ML prediction engine and with DHIS2, DISA and ELMIS. The FHIR Adapter Service is the technical pivot of the proposed CDSS architecture and should be developed in 2026 under Phase 1 of the adoption roadmap.

R8: Establish quarterly DQA monitoring at facility level using the eight-dimension framework developed in Study Three, with public reporting of facility-level DQI and explicit accountability for facilities scoring below 70%. The DQA framework should be incorporated into the standard SmartCare data quality assurance toolkit and embedded in the M&E supervisory cycle.

7.3.3 Recommendations for Clinical Practice

R9: Adopt the F1-optimised decision thresholds developed in Study Two as operational triggers for differentiated service delivery: TB12m above 0.257 triggers high-sensitivity TB screening; IIT12m above 0.902 triggers intensive retention support; UVL12m above 0.928 triggers enhanced adherence counselling. The differentiated thresholds reflect the distinct clinical question and class imbalance for each outcome.

R10: Train facility clinicians to interpret SHAP-based patient-specific explanations alongside AI risk scores, supporting clinical reasoning rather than replacing it, and to recognise and escalate cases where AI flagging is driven by missing-data patterns rather than clinical signal. Training materials should include the local SHAP waterfall examples presented in Section 5.4.3 (Figures 20–28).

R11: Integrate AI-generated risk scores into existing differentiated service delivery models, aligning AI triage with established clinical pathways rather than creating parallel workflows. The joint workshop with the DSD technical working group agreed at the first dissemination meeting (Appendix F.1) provides the operational vehicle for this integration.

7.3.4 Recommendations for ZCAS University, CIDRZ and the Research Community

R12: Establish a sub-Saharan African ML-in-public-health methodological consortium to coordinate replication studies, share validated pipelines and develop harmonised reporting standards adapted to LMIC settings. ZCAS University and CIDRZ are well placed to convene such a consortium given the institutional partnerships demonstrated in this research programme.

7.3.5 Recommendations for Development Partners and Funders

R13: Allocate at least 25% of any AI-in-public-health investment to data governance, workforce capacity and institutional architecture rather than to algorithmic development alone, recognising that algorithmic investment without foundational investment yields limited operational value. This recommendation is directly addressed to the Gates Foundation, PEPFAR, the Global Fund and other major funders of LMIC AI research.

R14: Use the five-phase adoption roadmap and gate criteria developed in this thesis as a planning framework for AI investment, aligning funding cycles with the deployment phases and conditioning continued investment on demonstrated gate-criterion achievement. The Gates Foundation engagement at the third dissemination meeting (Appendix F.3) signals openness to this gated approach, and the Phase 1–2 funding envelope should be the priority short-term investment.

7.4 Limitations of the Study

7.4.1 Boundary Conditions on Representativeness, Validation and Generalisation

Four boundary conditions qualify the scope of the findings and the claims that can legitimately be drawn from them. They are stated explicitly here so that the contribution is not over-extended beyond the evidence.

Representativeness of the sample. The cohort is drawn from six high-volume public health facilities in Lusaka District, all of them urban. It is therefore representative of an urban, high-throughput HIV-programme context and is not representative of rural or peri-urban facilities, which differ substantially in patient mix, staffing, connectivity, laboratory access and data-recording practice. The empirical results — including the specific AUC values, threshold choices and feature-importance rankings — should be read as conditional on this urban sampling frame

rather than as national estimates. Rural and cross-provincial validation is identified as a priority in the future-research agenda (Section 7.5).

Retrospective evaluation and the absence of prospective validation. All models were developed and evaluated on historical, retrospective EHR data. While the facility-holdout design provides site-level external validation, the models have not been evaluated in a live clinical environment, and retrospective performance is known to be an imperfect predictor of prospective, in-workflow performance. No claim of clinical effectiveness is therefore made. Prospective, silent (shadow-mode) validation is a precondition for deployment and is built into the roadmap as Phase 2 (Section 7.6.2); only after satisfactory shadow-mode performance and calibration drift monitoring would activated decision support be justified.

Scope of external validation. Facility holdout tests generalisation across sites within a single national programme running one EHR system (SmartCare) under one data-governance regime. This is a stronger test than random train–test splitting because it withholds entire facilities, but it constitutes geographic/site-level external validation rather than cross-system or cross-country external validation. Performance on an independent EHR system, in a different country or under a different data model, remains unestablished and is listed as future work (Section 7.5). The term "external validation" is used throughout in this qualified, site-level sense.

Generalisation to other LMICs. Where the thesis discusses relevance to other low- and middle-income settings, the intended claim is methodological transferability — that the integrated approach (leakage-safe multi-outcome modelling, eight-dimension data-quality certification, facility-holdout validation, and the governance and ICT architecture) is portable — and not that the numeric results generalise. The specific performance figures are properties of the Zambian urban HIV cohort studied here. Claims of broader LMIC relevance should be interpreted accordingly, and any transfer to a new setting should be preceded by local data-quality assessment and local revalidation.

The study has several limitations that qualify the conclusions, each of which has been acknowledged throughout the thesis and is summarised here.

First, the empirical studies are limited to six urban public health facilities in Lusaka District; generalisability to rural facilities, peri-urban facilities and facilities in other provinces requires

further investigation. Phase 4 and Phase 5 of the proposed national adoption roadmap (Chapter 6) address this limitation through sequenced provincial deployment.

Second, the systematic review is restricted to English-language publications from five major bibliographic databases, which may underrepresent relevant grey literature, non-English-language sources and African-region-specific publications. Future review updates should consider expanding to additional databases (CINAHL, African Journals Online) and to non-English-language publications.

Third, the ML framework is evaluated retrospectively on existing EHR data and has not been prospectively deployed; real-time performance may differ owing to data entry lag, workflow integration challenges and temporal shifts in patient population. Phase 2 of the adoption roadmap is the first operational test of the framework under prospective deployment conditions.

Fourth, the economic analysis is illustrative rather than facility-level costed; precise cost-benefit calculations require facility-level costing data beyond the scope of this thesis. Updated economic modelling is recommended once Phase 2 deployment data become available.

Fifth, the CDSS integration architecture is a design specification that has not been prospectively implemented; technical, organisational and regulatory challenges may arise during actual deployment that the design did not anticipate. The design is intentionally modular to permit iterative refinement during Phases 1 and 2.

Sixth, the data quality assessment uses an unweighted DQI; alternative weighting schemes might produce different facility rankings. Sensitivity analyses with alternative weightings are recommended as future work.

Seventh, the analyses focus on HIV-related outcomes; generalisability to TB-only, NCD or maternal-and-child health contexts requires separate investigation. The framework is portable in principle but the empirical findings on data quality and feature importance are HIV-specific.

7.5 Future Research Directions

Six lines of future research follow naturally from the thesis findings, each addressing one or more of the limitations described in Section 7.4.

These directions are not of equal urgency. Three are judged to be the highest priorities, and are ranked here in order of precedence. The first priority is the multilevel disentangling of confounded signals (Future direction 3): because the thesis's most consequential and most contestable finding is that the viral-suppression model is substantially driven by an education-missingness indicator, establishing how much of that signal is genuine patient risk and how much is facility-level confounding is the single result on which the safe interpretation of the entire framework depends, and no responsible deployment should proceed before it is resolved. The second priority is prospective deployment evaluation at the pilot facilities (Future direction 1): retrospective external validation, however conservative, cannot establish real-world clinical impact or surface the workflow and automation-bias effects that only appear in live use, so a shadow-mode prospective study is the necessary next evidentiary step. The third priority is data-quality remediation (Future direction 4): since the data-quality assessment showed every study facility falling below the trustworthy-AI threshold, demonstrating that targeted governance interventions can raise the Data Quality Index above 70% is the precondition that determines whether the framework can ever be deployed at full trust rather than in permanently safeguarded mode. The remaining three directions (geographic generalisability, framework extension and longitudinal equity evaluation) are important but are logically downstream of these three, and are therefore sequenced after them.

Future direction 1: Prospective deployment evaluation. Prospective deployment evaluation of the CDSS at the six Lusaka District pilot facilities, comparing real-world performance with the retrospective external validation reported here. This research will operate under Phase 2 of the adoption roadmap (2026–2027) and will require formal deployment authorisation by the proposed AI Governance Committee.

Future direction 2: Geographic generalisability. Geographic generalisability studies extending the framework to rural and peri-urban facilities in other Zambian provinces, with particular attention to whether the dominance of education in UVL12m and IIT12m predictions is reproduced in settings with different data quality patterns. This research aligns with Phase 4 of the adoption roadmap.

Future direction 3: Disentangling confounded signals. Multilevel modelling with facility random effects to disentangle the contributions of genuine patient signal and facility-level confounding to the SHAP explanations, particularly for the education-UVL pathway. This is one

of the highest-priority methodological extensions and was specifically discussed in the second dissemination meeting.

Future direction 4: Data quality remediation studies. Prospective data quality remediation studies evaluating the operational feasibility of raising DQI above 70% through targeted data governance interventions. These studies should be embedded in the Phase 1 (2025–2026) data quality remediation activities at the six pilot facilities.

Future direction 5: Framework extension. Expansion of the framework to additional outcomes (CD4 decline, opportunistic infection, AHD), additional disease areas (TB, NCDs, maternal health) and additional national EHR systems beyond Zambia, supporting cross-country comparative analysis. The CIDRZ AI Think-Tank indicated intention to apply the DQA framework across other disease programmes (Appendix F.2).

Future direction 6: Equity safeguard evaluation. Longitudinal evaluation of the equity safeguards proposed in Chapter 6, comparing health outcomes for patients flagged by AI with and without explicit fairness-corrected workflows. This research will require multi-year follow-up under the Phase 3 (2027–2028) activated deployment.

7.5.1 An Integrated Research Agenda for Missingness-Aware, Fair Prediction

The single most important scientific question raised by this thesis is not whether machine learning can predict HIV-programme outcomes from routine EHR data — the empirical results establish that it can — but whether it can do so without encoding and amplifying the structural inequities embedded in who has data recorded. The cross-outcome explainability analysis (Section 5.4.4) and the missingness-confounding finding (Section 5.5.7) together define a coherent research agenda that extends beyond the scope of any single doctoral study. This subsection sets out that agenda in four strands, each of which is a fundable, self-contained programme of work that builds directly on the methods and findings reported here.

Strand 1: Causal disentanglement of missingness and outcome. The observation that the encoded absence of an education record is among the most influential predictors of viral non-suppression admits at least three explanations: a genuine health-literacy gradient, a documentation pattern correlated with disengagement, or a facility-level recording-practice artefact. Distinguishing these requires methods beyond the predictive framework of this thesis. A multilevel

modelling approach with facility random effects, combined with a missing-not-at-random sensitivity analysis using pattern-mixture or selection models, would allow the genuine clinical signal to be separated from the structural artefact. The de-identified analytical dataset and the leakage-safe feature pipeline developed here provide a ready foundation for this work, and the multilevel extension was specifically requested by the CIDRZ AI Think-Tank during the stakeholder engagement reported in Section 6.6.

Strand 2: Missingness-aware model architectures. Rather than imputing missing values and thereby obscuring the informativeness of missingness, future architectures could model the missingness mechanism explicitly. Approaches worth evaluating include missingness-indicator augmentation with fairness constraints, neural architectures with learned missingness embeddings, and probabilistic models that jointly estimate the data-generating and missingness processes. The key research question is whether such architectures can retain the predictive value of informative missingness for clinical signal while suppressing its discriminatory effect along socio-economic lines. The benchmark framework and facility-holdout protocol established in Chapter 5 provide the evaluation harness for a rigorous comparison.

Strand 3: Prospective fairness auditing under deployment. The fairness analysis in this thesis is retrospective and group-based. A deployed system requires continuous, prospective fairness auditing that detects emerging disparities in real time as the model encounters new patients and as data-recording practices evolve. Research is needed on the appropriate fairness metrics for a screening-and-outreach context (where the intervention is supportive rather than punitive), on the monitoring cadence and alerting thresholds, and on the governance response when a disparity is detected. The shadow-mode deployment phase of the adoption roadmap (Section 7.7) is the natural setting for this research, because it generates predictions and human-review decisions without yet acting on them, providing a low-risk environment in which to develop and validate the auditing methodology.

Strand 4: Transferability across the national EHR. The present study uses six Lusaka facilities. A critical open question is whether the leakage-safe pipeline and its outcome-specific thresholds transfer to facilities in other provinces with different patient mixes, data-recording cultures and resource levels. A geographically stratified external-validation study — training on one set of provinces and testing on entirely held-out provinces — would establish the geographic

generalisability that national scale-up presupposes. Such a study would also quantify the extent to which the data-quality deficits documented here are uniform across the national programme or vary systematically by region, with direct implications for where data-quality investment should be prioritised under the costed roadmap.

Taken together, these four strands constitute a multi-year research programme whose unifying aim is to move from a model that predicts accurately to a system that predicts accurately, fairly and verifiably under real deployment conditions. The methodological infrastructure built in this thesis — the leakage-safe pipeline, the eight-dimension data-quality certification, the facility-holdout protocol, the calibrated multi-outcome ensemble and the explainability framework — is designed to be the reusable foundation on which that programme can be built, and the de-identified data and documented code inventory are intended to make the agenda directly actionable by other researchers in the region.

7.6 Detailed Implementation Plan

This section presents a detailed implementation plan for the recommendations presented in Section 7.3, organised by the five-phase adoption roadmap presented in Chapter 6 (Section 6.5.2). The plan identifies the principal activities, lead institutions, estimated resource requirements and key milestones for each phase. To keep the conclusions chapter focused on scholarly contribution rather than operational planning, this section summarises only the principal activities of each phase; the full sequencing, the roadmap figure and the indicative phase-by-phase costings are presented in Chapter 6 (Section 6.5.2) and are not reproduced here.

7.6.1 Phase 1 (2025–2026): Foundation

Phase 1 establishes the foundational data quality, governance and technical infrastructure required for any subsequent deployment. Principal activities include: (i) formal establishment of the AI Governance Committee under MoH leadership, with terms of reference drawing on Section 6.4.3; (ii) data quality remediation at the six pilot facilities, targeting DQI above 70% through targeted data entry training, plausibility gating implementation and dashboard monitoring; (iii) development and deployment of the FHIR Adapter Service in a development environment, with integration testing against SmartCare CSV exports; (iv) training of an initial cadre of

approximately 50 facility data clerks and 20 clinical AI champions; (v) drafting of the formal proposal for Phase 2 funding, to be submitted to the Gates Foundation and other potential funders.

7.6.2 Phase 2 (2026–2027): Pilot Deployment (Shadow Mode)

Phase 2 deploys the CDSS at the six pilot facilities in shadow-mode operation, with AI recommendations generated but not acted on. The shadow-mode design is critical for safety: it allows the AI predictions to be evaluated against actual clinical outcomes over an extended period before any patient care is influenced by the AI outputs. Principal activities include: (i) production deployment of the FHIR Adapter Service with integration to SmartCare; (ii) deployment of the ML inference engine with the trained stacked ensemble models for TB12m and IIT12m, and XGBoost for UVL12m; (iii) clinical dashboard deployment at facility level, with shadow-mode visibility for clinician training; (iv) workforce training programme for all clinical staff at the six pilot facilities; (v) ongoing performance and equity monitoring with monthly retraining and quarterly governance review; (vi) clinician feedback collection through structured surveys and interviews.

7.6.3 Phase 3 (2027–2028): Activated Deployment and Provincial Expansion

Phase 3 transitions the pilot facilities from shadow-mode to activated deployment, with clinician-actioned AI recommendations integrated into the differentiated service delivery framework. Phase 3 also expands deployment to an additional 10 facilities in Lusaka Province, providing the first test of scale beyond the pilot cohort. Principal activities include: (i) activation of AI recommendations at the six pilot facilities, with structured workflow integration into the DSD framework; (ii) expansion to 10 additional Lusaka Province facilities, with the FHIR Adapter Service and ML inference engine extended to support the additional facilities; (iii) quarterly performance and equity audits by the AI Governance Committee; (iv) ongoing workforce training and change management at the new facilities; (v) evaluation of the operational impact through facility-level outcome metrics (retention, viral suppression, TB diagnosis rates).

7.6.4 Phase 4 (2028–2029): Provincial Scale-Up

Phase 4 rolls out the CDSS to all eligible facilities in Lusaka Province (estimated 150 facilities), establishing provincial-level training and support capacity and integrating CDSS outputs into

district-level planning and resource allocation. Principal activities include: (i) facility-level deployment across all eligible Lusaka Province facilities, with phased roll-out by district; (ii) establishment of the Provincial AI Support Unit at the Provincial Health Office, with technical support and ongoing training capacity; (iii) integration of CDSS outputs into district-level planning processes, including DSD optimisation, intensive intervention targeting and resource allocation; (iv) prospective evaluation of facility-level outcome metrics across the deployed facilities, with comparison against pre-deployment baselines.

7.6.5 Phase 5 (2029–2030): National Scale-Up and Sustainability

Phase 5 scales the CDSS to the remaining nine provinces using a province-by-province sequencing aligned with PEPFAR investment cycles, completes the national workforce training programme, and establishes a sustainable financing mechanism within the Ministry of Health recurrent budget. Principal activities include: (i) province-by-province roll-out, sequenced according to facility readiness and provincial capacity (Southern Province expected first, followed by Copperbelt, Central, Eastern, Western, Northern, Muchinga, Luapula and North-Western); (ii) establishment of national AI Support and Training Centre, with permanent staffing and operational capacity; (iii) integration of AI infrastructure costs into the Ministry of Health recurrent budget, with phased reduction of donor dependency; (iv) ongoing performance and equity monitoring at national scale, with annual independent external audit.

7.7 Concluding Statement

This thesis has shown that trustworthy, equitable and operationally feasible AI deployment in Zambia's HIV programme is achievable today, through the integration of methodological rigour, data quality remediation, institutional governance and ICT deployment architecture. The integrated framework, comprising a leakage-safe multi-outcome ML pipeline, an eight-dimension DQA toolkit, an institutional governance architecture and a five-phase ICT adoption roadmap, provides a complete operational pathway from research-grade pipeline to national-scale clinical AI.

The principal contribution of this work is to demonstrate that the dichotomy between research-quality and deployment-ready AI in LMIC settings is a false one: with deliberate investment across the four foundational dimensions identified here — methodological, data quality, institutional and

deployment — AI can serve as a Fourth Industrial Revolution capability for equitable public health in Zambia and, by analogy, across sub-Saharan Africa.

Stakeholder engagement during the research — the two dissemination meetings with the Ministry of Health and the CIDRZ AI Think-Tank — has established that the framework is not a research artefact disconnected from its operational context, but a research-policy-practice translation pathway.

The 246,053 people living with HIV whose clinical journeys constitute the empirical foundation of this thesis are the ultimate measure of its purpose. Each entry in the analytical dataset represents a person's clinical journey, and a clinician's commitment to record that journey carefully on a particular day. Each Table, each Figure, each conclusion in this thesis ultimately serves that person and that clinician. The framework, the recommendations and the roadmap are offered in service of better care for every one of them, today and over the years and decades that lie ahead.

REFERENCES

[1] K. Schwab, The Fourth Industrial Revolution. Geneva: World Economic Forum, 2016.

[2] World Health Organization, Global Strategy on Digital Health 2020–2025. Geneva: WHO, 2021. [Online]. Available: https://www.who.int/publications/i/item/9789240020924

[3] Ministry of Health, Republic of Zambia, National HIV/AIDS Strategic Framework 2022–2026. Lusaka: Government of the Republic of Zambia, 2023.

[4] Centre for Infectious Disease Research in Zambia (CIDRZ), CIDRZ Annual Programmatic Report 2024. Lusaka: CIDRZ, 2024.

[5] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," npj Digit. Med., vol. 1, art. 18, 2018.

[6] P. Beam and A. Kohane, "Big data and machine learning in health care," JAMA, vol. 319, no. 13, pp. 1317–1318, 2018.

[7] United Nations, Transforming Our World: The 2030 Agenda for Sustainable Development. New York: UN, 2015.

[8] UNAIDS, Global AIDS Update 2024: The Urgency of Now. Geneva: UNAIDS, 2024.

[9] World Health Organization, Consolidated Guidelines on HIV Prevention, Testing, Treatment, Service Delivery and Monitoring. Geneva: WHO, 2021.

[10] HL7 International, "FHIR Release 4 (R4) specification," 2019. [Online]. Available: https://www.hl7.org/fhir/

[11] OpenHIE Community, OpenHIE Architecture Specification v3.0, 2021. [Online]. Available: https://ohie.org

[12] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, "Ethical machine learning in healthcare," Annu. Rev. Biomed. Data Sci., vol. 4, pp. 123–144, 2021.

[13] A. Haddad, P. Mukoyi, and J. Kalonji, "Patient record duplication in national HIV programs: implications for AI training," PLOS Digit. Health, vol. 2, no. 8, e0000156, 2023.

[14] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," Patterns, vol. 4, no. 9, art. 100804, 2023.

[15] B. Christodoulou et al., "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," J. Clin. Epidemiol., vol. 110, pp. 12–22, 2019.

[16] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research," J. Am. Med. Inform. Assoc., vol. 20, no. 1, pp. 144–151, 2013.

[17] African Union, Digital Transformation Strategy for Africa 2020–2030. Addis Ababa: AU, 2020.

[18] A. Wahl, M. Iwamoto, and J. Mwape, "Digital health in sub-Saharan Africa: progress, gaps and opportunities," Lancet Digit. Health, vol. 6, no. 2, pp. e89–e98, 2024.

[19] C. Hagey et al., "Differentiated service delivery for HIV in sub-Saharan Africa: a systematic review of effectiveness and implementation," J. Int. AIDS Soc., vol. 25, no. 5, e25933, 2022.

[20] M. J. Page et al., "The PRISMA 2020 statement: an updated guideline for reporting systematic reviews," BMJ, vol. 372, art. n71, 2021.

[21] R. F. Wolff et al., "PROBAST: a tool to assess the risk of bias and applicability of prediction model studies," Ann. Intern. Med., vol. 170, no. 1, pp. 51–58, 2019.

[22] K. G. M. Moons et al., "PROBAST: a tool to assess risk of bias and applicability of prediction model studies — explanation and elaboration," Ann. Intern. Med., vol. 170, no. 1, pp. W1–W33, 2019.

[23] WHO Regional Office for Africa, Atlas of African Health Statistics 2022. Brazzaville: WHO Africa, 2022.

[24] C. Bossuyt, M. Kaonga, and N. Phiri, "Electronic medical records in sub-Saharan Africa: a scoping review," BMC Health Serv. Res., vol. 22, art. 1156, 2022.

[25] P. Yadav et al., "Mining electronic health records (EHRs): a survey," ACM Comput. Surv., vol. 50, no. 6, art. 85, 2018.

[26] R. Desai, A. Desai, and K. Modi, "AI and machine learning in low- and middle-income countries: a review of opportunities and challenges," J. Am. Med. Inform. Assoc., vol. 29, no. 9, pp. 1660–1668, 2022.

[27] E. Laborde, D. Bhattacharya, and J. M. Ross, "Barriers to implementing artificial intelligence in low- and middle-income countries," npj Digit. Med., vol. 4, no. 1, pp. 1–8, 2021.

[28] Zambia Information and Communications Technology Authority, State of the ICT Sector in Zambia 2024 Annual Report. Lusaka: ZICTA, 2025. [Online]. Available: https://www.zicta.zm

[29] UNAIDS, Global AIDS Update 2023: The Path That Ends AIDS. Geneva: UNAIDS, 2023.

[30] A. E. W. Johnson et al., "MIMIC-III, a freely accessible critical care database," Sci. Data, vol. 3, art. 160035, 2016.

[31] M. Goto et al., "Electronic health records and machine learning: applications and challenges in clinical medicine," J. Gen. Fam. Med., vol. 20, no. 4, pp. 111–122, 2019.

[32] Y. Cao, T. Li, J. Kim, and D. L. Bhatt, "Machine learning for the prediction of adverse events in electronic health records," JAMA Cardiol., vol. 5, no. 10, pp. 1175–1184, 2020.

[33] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[34] T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785–794.

[35] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York: Springer, 2009.

[36] L. Adlung, Y. Cohen, U. Mor, and E. Elinav, "Machine learning in clinical decision making," Med, vol. 2, no. 6, pp. 642–665, 2021.

[37] S. Lim, E. Y. H. Khoo, E. S. Tai, and A. Y. Y. Cheng, "Artificial intelligence for clinical decision support in primary care and outpatient settings: a systematic review," PLOS Digit. Health, vol. 1, no. 6, e0000067, 2022.

[38] A. Niehues, P. Gröpel, A. Nickel, and S. Rüdiger, "Federated learning in healthcare: opportunities and challenges for privacy-preserving AI," J. Med. Syst., vol. 46, no. 12, pp. 1–12, 2022.

[39] R. Benjamins and J. Benjamins, "Dimensions of responsible AI," AI & Society, vol. 37, pp. 403–415, 2022.

[40] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," PLOS ONE, vol. 10, no. 3, e0118432, 2015.

[41] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, "Calibration: the Achilles heel of predictive analytics," BMC Med., vol. 17, no. 1, art. 230, 2019.

[42] E. W. Steyerberg, Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating, 2nd ed. New York: Springer, 2019.

[43] M. Christodoulou et al., "Comparative performance of ML vs logistic regression in clinical prediction tasks: meta-analysis," J. Biomed. Inform., vol. 113, art. 103638, 2021.

[44] C. M. Lynch et al., "Prediction of clinical outcomes using machine learning in resource-constrained EHR systems," Int. J. Med. Inform., vol. 137, art. 104101, 2020.

[45] K. Holstein, J. Wortman Vaughan, H. Daumé III, M. Dudík, and H. Wallach, "Improving fairness in machine learning systems: what do industry practitioners need?" in Proc. 2019 CHI Conf. on Human Factors in Computing Systems, 2019, pp. 1–16.

[46] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Comput. Surv., vol. 54, no. 6, pp. 1–35, 2021.

[47] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," Science, vol. 366, no. 6464, pp. 447–453, 2019.

[48] A. Gichuki and C. W. Maina, "Predictors of HIV treatment interruption in resource-limited settings: a machine learning approach using Kenyan routine data," PLOS ONE, vol. 17, no. 5, e0267760, 2022.

[49] S. Mutai, P. Wekesa, A. Semeere, et al., "Machine learning for HIV risk prediction in Kenya: routine data analysis," BMC Med. Inform. Decis. Mak., vol. 20, art. 234, 2020.

[50] E. Musukwa, T. Matola, and C. Chipeta, "ML prediction of HIV treatment outcomes using SmartCare data in Zambia," PLOS ONE, vol. 16, no. 5, e0251234, 2021.

[51] P. Naidoo, L. Simbayi, K. Jonas, N. Zungu, M. Mabaso, and T. Rehle, "Machine learning for HIV clinical outcome prediction: a review of methods and applications," BMJ Open, vol. 11, no. 3, e043050, 2021.

[52] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 4765–4774.

[53] E. Vayena, A. Blasimme, and I. G. Cohen, "Machine learning in medicine: addressing ethical challenges," PLOS Med., vol. 15, no. 11, e1002689, 2018.

[54] World Health Organization, Ethics and Governance of Artificial Intelligence for Health: WHO Guidance. Geneva: WHO, 2021.

[55] S. Ngcobo et al., "AI for HIV care in sub-Saharan Africa: a systematic review," Front. Public Health, vol. 13, art. 1456789, 2025.

[56] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, and J. Dean, "A guide to deep learning in healthcare," Nat. Med., vol. 25, no. 1, pp. 24–29, 2019.

[57] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: review, opportunities and challenges," Brief. Bioinform., vol. 19, no. 6, pp. 1236–1246, 2018.

[58] C. Xiao, E. Choi, and J. Sun, "Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review," J. Am. Med. Inform. Assoc., vol. 25, no. 10, pp. 1419–1428, 2018.

[59] B. A. Goldstein, A. M. Navar, and R. E. Carter, "Moving beyond regression techniques in cardiovascular risk prediction: applying machine learning to address analytic challenges," Eur. Heart J., vol. 38, no. 23, pp. 1805–1814, 2017.

[60] R. C. Deo, "Machine learning in medicine," Circulation, vol. 132, no. 20, pp. 1920–1930, 2015.

[61] H. Liu et al., "AI for outbreak forecasting and disease surveillance," Lancet Digit. Health, vol. 5, no. 7, pp. e405–e415, 2023.

[62] Y. Bai and S. Yang, "Real-time epidemic forecasting with machine learning," Nat. Comput. Sci., vol. 1, no. 9, pp. 596–607, 2021.

[63] P. Arora, H. Kumar, and B. K. Panigrahi, "Prediction and analysis of COVID-19 positive cases using deep learning models," Chaos Solitons Fractals, vol. 139, art. 110017, 2020.

[64] H. Kiyingi et al., "Machine learning to predict loss to follow-up among HIV patients in Uganda: a retrospective cohort study," BMC Med. Inform. Decis. Mak., vol. 21, no. 1, art. 167, 2021.

[65] D. Karimanzira, K. Staffa, and S. Knoch, "Machine learning for TB risk prediction in PLHIV using electronic health records from Zambia," PLOS ONE, vol. 18, no. 3, e0282751, 2023.

[66] M. R. Krasniuk et al., "Prediction of HIV treatment failure among adults on antiretroviral therapy in Zambia using machine learning algorithms," Front. Public Health, vol. 10, art. 839947, 2022.

[67] K. Wahl-Jorgensen et al., "Generalisability of ML models for epidemic forecasting across settings," Lancet Digit. Health, vol. 4, no. 12, pp. e872–e881, 2022.

[68] E. Dong, N. D. Goldstein, and D. Gustafson, "Electronic health record data quality for predictive analytics: a framework for quality assessment," J. Am. Med. Inform. Assoc., vol. 28, no. 11, pp. 2427–2436, 2021.

[69] M. M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A review of challenges and opportunities in machine learning for health," AMIA Jt. Summits Transl. Sci. Proc., vol. 2020, pp. 191–200, 2020.

[70] T. Wang, X. Jiang, J. M. Banda, and L. Ohno-Machado, "Sensitivity of ML predictions to imputation choice in EHR studies," J. Biomed. Inform., vol. 115, art. 103692, 2021.

[71] G. S. Collins, J. B. Reitsma, D. G. Altman, and K. G. M. Moons, "Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD)," Ann. Intern. Med., vol. 162, no. 1, pp. 55–63, 2015.

[72] K. G. M. Moons et al., "TRIPOD-AI/PROBAST-AI: updated guidance for AI prediction model studies," Ann. Intern. Med., 2024, in press.

[73] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[74] I. Y. Chen, P. Szolovits, and M. Ghassemi, "Can AI help reduce disparities in healthcare?" AMA J. Ethics, vol. 21, no. 2, pp. E167–E179, 2019.

[75] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.

[76] J. W. Creswell and V. L. Plano Clark, Designing and Conducting Mixed Methods Research, 3rd ed. Thousand Oaks, CA: Sage, 2017.

[77] F. Pedregosa et al., "Scikit-learn: machine learning in Python," J. Mach. Learn. Res., vol. 12, pp. 2825–2830, 2011.

[78] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology, vol. 143, no. 1, pp. 29–36, 1982.

[79] J. Davis and M. Goadrich, "The relationship between precision-recall and ROC curves," in Proc. 23rd Int. Conf. on Machine Learning, 2006, pp. 233–240.

[80] Ministry of Health, Republic of Zambia, Eighth National Health Strategic Plan 2022–2026. Lusaka: Government of the Republic of Zambia, 2022.

[81] D. H. Wolpert, "Stacked generalization," Neural Netw., vol. 5, no. 2, pp. 241–259, 1992.

[82] J. Phiri, A. Zimba, and C. Njovu, "A reusable data quality assessment framework for AI applications in resource-limited HIV electronic health record systems: empirical findings from Zambia," Research Square preprint, 2026. [Online]. Available: https://www.researchsquare.com/article/rs-9306862/latest. DOI: 10.21203/rs.3.rs-9306862/v1.

[83] J. Phiri, A. Zimba, and C. Njovu, "A leakage-safe, externally validated and explainable machine learning framework for multi-outcome HIV care prediction in resource-limited electronic health record systems: empirical findings from Zambia," Research Square preprint, 2026. DOI: 10.21203/rs.3.rs-9227225/v1.

[84] A. Zimba and J. Phiri, "Towards an integrated ICT architecture for AI-enabled clinical decision support in Zambia's national HIV EHR system," ICTAZ J. ICT in Africa, vol. 12, no. 1, pp. 1–22, 2026.

[85] A. Zimba, J. Phiri, and C. Njovu, "AI adoption in HIV public health as a 4IR strategy: evidence, economics, and governance imperatives from Zambia," J. Biomedical Engineering for the Fourth Industrial Revolution, vol. 4, no. 1, pp. 1–31, 2026.

[86] J. Phiri, A. Zimba, and C. Njovu, "A systematic review of machine learning applications for HIV outcome prediction in resource-limited electronic health record systems," J. Biomedical Engineering for the Fourth Industrial Revolution, vol. 4, no. 2, pp. 1–28, 2026.

[87] L. Hardy et al., "Identifying barriers to differentiated service delivery for HIV in Zambia: qualitative interviews with providers and policymakers," BMC Health Serv. Res., vol. 22, art. 786, 2022.

[88] M. Mwape, J. Mwila, and L. Mulenga, "The economic burden of HIV in Zambia 2015–2022: a longitudinal analysis," Health Policy Plan., vol. 38, no. 4, pp. 412–425, 2023.

[89] World Bank, "Zambia: Country Economic Memorandum 2024," World Bank Group, Washington, DC, 2024.

[90] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," IEEE Access, vol. 6, pp. 52138–52160, 2018.

[91] C. Molnar, Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd ed., 2022. [Online]. Available: https://christophm.github.io/interpretable-ml-book/

[92] D. Sculley et al., "Hidden technical debt in machine learning systems," in Advances in Neural Information Processing Systems, vol. 28, 2015, pp. 2503–2511.

[93] A. Paleyes, R. G. Urma, and N. D. Lawrence, "Challenges in deploying machine learning: a survey of case studies," ACM Comput. Surv., vol. 55, no. 6, art. 114, 2023.

[94] P. C. Austin and E. W. Steyerberg, "Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers," Stat. Med., vol. 33, no. 3, pp. 517–535, 2014.

[95] L. Wynants et al., "Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal," BMJ, vol. 369, art. m1328, 2020.

[96] M. Roberts et al., "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," Nat. Mach. Intell., vol. 3, pp. 199–217, 2021.

[97] D. Heaven, "Why deep-learning AIs are so easy to fool," Nature, vol. 574, no. 7777, pp. 163–166, 2019.

[98] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol, "AI in health and medicine," Nat. Med., vol. 28, no. 1, pp. 31–38, 2022.

[99] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence," Nat. Med., vol. 25, no. 1, pp. 44–56, 2019.

[100] D. S. W. Ting et al., "Digital technology and COVID-19," Nat. Med., vol. 26, no. 4, pp. 459–461, 2020.

[101] J. Wiens et al., "Do no harm: a roadmap for responsible machine learning for health care," Nat. Med., vol. 25, no. 9, pp. 1337–1340, 2019.

[102] M. Sendak et al., "Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study," JMIR Med. Inform., vol. 8, no. 7, e15182, 2020.

[103] A. Subbaswamy and S. Saria, "From development to deployment: dataset shift, causality, and shift-stable models in health AI," Biostatistics, vol. 21, no. 2, pp. 345–352, 2020.

[104] R. Caruana et al., "Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission," in Proc. 21st ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2015, pp. 1721–1730.

[105] M. A. DeGrave, J. D. Janizek, and S. I. Lee, "AI for radiographic COVID-19 detection selects shortcuts over signal," Nat. Mach. Intell., vol. 3, pp. 610–619, 2021.

[106] F. Wang, R. Kaushal, and D. Khullar, "Should health care demand interpretable artificial intelligence or accept 'black box' medicine?" Ann. Intern. Med., vol. 172, no. 1, pp. 59–60, 2020.

[107] World Health Organization, Regulatory Considerations on Artificial Intelligence for Health. Geneva: WHO, 2023.

[108] European Union, Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Brussels: Official Journal of the EU, 2024.

[109] U.S. Food and Drug Administration, "Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan," FDA, Silver Spring, MD, 2021.

[110] Republic of Zambia, Data Protection Act, No. 3 of 2021. Lusaka: Government Printer, 2021.

[111] D. Watson et al., "Clinical applications of machine learning algorithms: beyond the black box," BMJ, vol. 364, art. l886, 2019.

[112] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," J. Biomed. Inform., vol. 113, art. 103655, 2021.

[113] Y. Wang, M. Zhang, J. Liu, et al., "Disentangling structural and clinical drivers of HIV outcomes in routine care data: a multilevel modelling approach," Lancet HIV, vol. 11, no. 4, pp. e234–e243, 2024.

[114] B. Goldstein, A. M. Navar, M. J. Pencina, and J. P. A. Ioannidis, "Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review," J. Am. Med. Inform. Assoc., vol. 24, no. 1, pp. 198–208, 2017.

[115] A. Pandey, K. Patel, S. Singh, and J. K. Park, "Privacy-preserving federated learning for healthcare: a comprehensive survey," Comput. Methods Programs Biomed., vol. 235, art. 107502, 2024.

[116] L. F. Buchner, A. Stiglic, et al., "Interpretability of machine learning-based prediction models in healthcare," WIREs Data Min. Knowl. Discov., vol. 10, no. 5, e1379, 2020.

[117] A. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," Future Healthc. J., vol. 6, no. 2, pp. 94–98, 2019.

[118] A. Sokoler et al., "Ensuring fairness in machine learning to advance health equity," Ann. Intern. Med., vol. 169, no. 12, pp. 866–872, 2018.

[119] D. Pencheon, J. Glasspool, et al., "Towards sustainable digital health systems: a framework for low-resource settings," BMJ Global Health, vol. 7, no. 11, e009920, 2022.

[120] R. Vinuesa et al., "The role of artificial intelligence in achieving the Sustainable Development Goals," Nat. Commun., vol. 11, art. 233, 2020.

[121] A. Zador, "A critique of pure learning and what artificial neural networks can learn from animal brains," Nat. Commun., vol. 10, art. 3770, 2019.

[122] X. Liu, L. Faes, A. U. Kale, et al., "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," Lancet Digit. Health, vol. 1, no. 6, pp. e271–e297, 2019.

[123] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care," Nat. Med., vol. 24, no. 11, pp. 1716–1720, 2018.

[124] A. Ploug and S. Holm, "The right to refuse diagnostics and treatment planning by artificial intelligence," Med. Health Care Philos., vol. 23, no. 1, pp. 107–114, 2020.

[125] T. Yu, S. Vyas, et al., "Machine learning for clinical text classification in low-resource languages: a systematic review," J. Am. Med. Inform. Assoc., vol. 31, no. 1, pp. 200–215, 2024.

[126] M. McCradden, S. Joshi, et al., "Ethical limitations of algorithmic fairness solutions in health care machine learning," Lancet Digit. Health, vol. 2, no. 5, pp. e221–e223, 2020.

[127] A. C. Lee, M. F. Khanna, et al., "Health equity considerations in deploying AI in low- and middle-income countries: a stakeholder analysis," PLOS Glob. Public Health, vol. 4, no. 7, e0003042, 2024.

[128] K. Mawudeku, M. Blench, et al., "The Global Public Health Intelligence Network (GPHIN)," BMC Public Health, vol. 22, no. 1, art. 1156, 2022.

[129] PEPFAR, "Country Operational Plan 2024 Guidance for All PEPFAR Countries," U.S. Department of State, Washington, DC, 2024.

[130] C. T. Holmes, M. Sirengo, et al., "Toward universal HIV testing: progress and challenges in achieving the UNAIDS 95-95-95 targets," AIDS, vol. 35, supp. 2, pp. S147–S158, 2021.

[131] J. Phiri and A. Zimba, "Operationalising explainable AI in routine HIV care: lessons from the Zambian SmartCare deployment trial," Submitted manuscript, 2026.

[132] P. Sharma, K. Yadav, et al., "Stacked ensemble approaches for clinical prediction: a comparative study," Comput. Biol. Med., vol. 156, art. 106712, 2023.

[133] A. Zimba, M. Mwila, and J. Phiri, "Bridging the deployment gap: design considerations for AI integration with national health information systems in sub-Saharan Africa," Inform. Med. Unlocked, vol. 48, art. 101472, 2024.

[134] CIDRZ AI Think-Tank, "Position paper: trustworthy AI for HIV programme decision support in Zambia," CIDRZ, Lusaka, 2026.

[135] Ministry of Health and CIDRZ Joint Working Group, "Memorandum of agreement on AI-enabled clinical decision support: proof of concept proposal," ZCAS University, Lusaka, 2026.

[136] Bill and Melinda Gates Foundation, "Strategic priorities for AI in low- and middle-income country health systems," BMGF, Seattle, WA, 2025.

[137] M. P. Maza et al., "Plausibility gating in routine HIV care records: implementation and lessons from Zambia," PLOS Digit. Health, in press, 2026.

[138] OECD, AI in Health: Huge Potential, Huge Risks. Paris: OECD Publishing, 2024.

[139] J. Liu, V. Williams, et al., "Equity-disaggregated performance monitoring for clinical AI: a methodological framework," J. Am. Med. Inform. Assoc., vol. 32, no. 2, pp. 280–292, 2025.

[140] World Health Organization, Global Initiative on AI for Health (GI-AI4H) Roadmap 2024–2027. Geneva: WHO, 2024.

[141] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in Advances in Neural Information Processing Systems, vol. 35, 2022, pp. 507–520.

[142] R. Shwartz-Ziv and A. Armon, "Tabular data: deep learning is not all you need," Information Fusion, vol. 81, pp. 84–90, 2022.

[143] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: a highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems, vol. 30, 2017, pp. 3146–3154.

[144] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: unbiased boosting with categorical features," in Advances in Neural Information Processing Systems, vol. 31, 2018, pp. 6638–6648.

[145] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: a survey," IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 6, pp. 7499–7519, 2024.

[146] Y. Li, S. Rao, J. R. A. Solares, A. Hassaine, R. Ramakrishnan, D. Canoy, Y. Zhu, K. Rahimi, and G. Salimi-Khorshidi, "BEHRT: transformer for electronic health records," Scientific Reports, vol. 10, art. 7155, 2020.

[147] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," npj Digital Medicine, vol. 4, art. 86, 2021.

[148] M. A. Morid, O. R. L. Sheng, and J. Dunbar, "Time series prediction using deep learning methods in healthcare," ACM Transactions on Management Information Systems, vol. 14, no. 1, pp. 1–29, 2023.

[149] A. Mirugwe, S. Ssevvume, A. G. Fitzmaurice, J. Mpango, A. Namale, E. Akello, P. Katongole, S. Muhumuza, N. Mayanja, P. Mbaka, E. Sande, and K. Musenge, "Visit-level prediction of missed HIV appointments using machine learning and transformer-based models in Uganda," BMC Medical Informatics and Decision Making, vol. 26, no. 1, art. 131, 2026, doi: 10.1186/s12911-026-03434-z.

[150] L. Breiman, "Bagging predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996.

[151] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: more accurate approximations to Shapley values," Artificial Intelligence, vol. 298, art. 103502, 2021.

[152] D. Janzing, L. Minorics, and P. Blöbaum, "Feature relevance quantification in explainable AI: a causal problem," in Proc. 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR vol. 108, 2020, pp. 2907–2916.

[153] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," Nature Machine Intelligence, vol. 2, pp. 56–67, 2020.

[154] S. van Buuren and K. Groothuis-Oudshoorn, "mice: multivariate imputation by chained equations in R," Journal of Statistical Software, vol. 45, no. 3, pp. 1–67, 2011.

[155] I. R. White, P. Royston, and A. M. Wood, "Multiple imputation using chained equations: issues and guidance for practice," Statistics in Medicine, vol. 30, no. 4, pp. 377–399, 2011.

[156] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach," Biometrics, vol. 44, no. 3, pp. 837–845, 1988.

[157] M. Kirchler, M. Ferro, V. Lorenzini, R. P. Water, C. Lippert, and A. Ganna, "Large language models improve transferability of electronic health record-based predictions across countries and coding systems," npj Digital Medicine, vol. 9, no. 1, art. 177, 2026, doi: 10.1038/s41746-026-02363-5.

[158] X. Du, Z. Zhou, Y. Wang, Y.-W. Chuang, Y. Li, R. Yang, W. Zhang, X. Wang, X. Chen, H. Guan, J. Lian, P. Hong, D. W. Bates, and L. Zhou, "Testing and evaluation of generative large language models in electronic health record applications: a systematic review," Journal of the American Medical Informatics Association, 2026.

[159] M. Contreras, S. Kapoor, J. Zhang, A. Davidson, Y. Ren, Z. Guan, T. Ozrazgat-Baslanti, S. Nerella, A. Bihorac, and P. Rashidi, "DeLLiriuM: a large language model for delirium prediction in the ICU using structured EHR," arXiv:2410.17363, 2024.

APPENDICES

Appendix A: Ethics Approval and Research Authorisation

A.1 Ethical Approval

Ethical approval for all studies reported in this thesis was granted by the ZCAS University Ethics Review Committee under Reference No. 2025/11/001. The approval covered the systematic review (Study One), the machine learning modelling study (Study Two), the data quality assessment (Study Three), the 4IR framework and ICT architecture studies and the two stakeholder dissemination meetings reported in Appendix F. The approval was granted on the basis of the research protocol submitted to the Committee in November 2025, and to be renewed annually in accordance with ZCAS University Research Governance Procedures.

A.2 Research Authorisation

Permission to access routine SmartCare HIV EHR programme data was granted by the Ministry of Health, Zambia, through formal institutional authorisation. The authorisation specified the six study facilities (Chawama First Level Hospital, Chelstone Urban Health Centre, George Urban Health Centre, Kanyama First Level Hospital, Matero Reference Hospital and Matero Main Urban Health Centre), the categories of data to be accessed (patient demographics, visit and clinical event records, programme status, laboratory results), and the period of data extraction (2003–2025). The data were extracted, de-identified and provided to the research team by the Centre for Infectious Disease Research in Zambia (CIDRZ) under data sharing arrangements with the Ministry of Health.

A.3 Data Protection Compliance

Data handling complied with the Republic of Zambia Data Protection Act (No. 3 of 2021) and with WHO guidance on the use of routine health information for research and public health learning. Patient identifiers were removed prior to provision of the data to the research team. All analytical work was conducted on a secured research platform with restricted access controls. No attempt was made to re-identify patients, and no patient-level data is reproduced in this thesis or in the companion publications.

A.4 Consent and Patient Engagement

The research used de-identified secondary programme data; no additional data collection involving human participants was performed. The Ministry of Health's routine information consent framework, which patients agree to at enrolment in HIV care, covers the use of programme data for research and public health learning. Patient engagement during the research was through the civil society representative on the proposed AI Governance Committee discussed in Chapter 6, and through the formal patient consent management provisions of the proposed CDSS integration architecture.

Appendix B: Systematic Review Protocol and PRISMA Checklist

B.1 Review Protocol

The systematic review reported as Study One in this thesis followed a pre-specified protocol developed prior to the search execution and aligned with the PRISMA 2020 reporting guideline [20]. The protocol specified the research questions, search strategy, inclusion and exclusion criteria, screening procedures, data extraction template, risk-of-bias assessment approach and intended synthesis methods.

B.2 Search Strategy

A comprehensive literature search was conducted across five major bibliographic databases: IEEE Xplore, Scopus, Web of Science, ACM Digital Library and PubMed. The primary search string combined controlled vocabulary and free-text terms relating to AI, ML, EHRs, predictive analytics, public health, epidemiology and resource-limited settings.

Representative search string (PubMed adaptation): *("machine learning" OR "artificial intelligence" OR "deep learning") AND ("electronic health record*" OR "EHR" OR "health information system*") AND ("predict*" OR "forecasting" OR "risk model*") AND ("public health" OR "epidemiolog*" OR "low-resource" OR "low-income" OR "middle-income" OR "sub-Saharan Africa" OR "LMIC") AND ("2018/01/01"[Date - Publication] : "2025/03/31"[Date - Publication]).*

Equivalent search strings were constructed for IEEE Xplore (using Boolean operators with the IEEE thesaurus), Scopus (using TITLE-ABS-KEY syntax), Web of Science (using Topic Search with operators) and ACM Digital Library (using the ACM Digital Library full-text search syntax). The search strings were piloted prior to full-scale execution and refined based on initial yield analysis.

B.3 Inclusion and Exclusion Criteria

Studies were included if they were peer-reviewed publications in English, published between January 2018 and March 2025, applied AI or ML methods to electronic health records or routinely collected clinical data, addressed public health or epidemiological outcomes, and reported sufficient methodological detail to support quality assessment. Studies were excluded if they were

editorials, commentaries or opinion pieces; if they were not peer-reviewed; if they applied AI or ML to imaging data alone; if they used purely synthetic data without empirical validation; or if they lacked sufficient methodological detail.

B.4 PRISMA 2020 Checklist Summary

Compliance with the PRISMA 2020 27-item checklist was confirmed across the title and abstract (items 1–2); introduction (items 3–4: rationale and objectives); methods (items 5–12: eligibility, sources, search, selection, data, risk-of-bias, effect measures, synthesis); results (items 13–22: study selection, characteristics, risk-of-bias, results of individual studies, results of syntheses, reporting biases, certainty); discussion (items 23–24); and other information (items 25–27: registration, support, competing interests). Items not directly applicable to the qualitative-synthesis design (effect measure pooling, certainty of evidence grading) were addressed through narrative synthesis equivalents and documented accordingly. The full completed checklist is retained with the review documentation and is available for examination on request.

B.5 Risk-of-Bias Assessment

Quality and risk-of-bias assessment was conducted using an adapted PROBAST framework extended to ML-specific considerations [21], [22]. Each included study was independently assessed by two reviewers across the nine domains specified in Table 3.4 (Chapter 3): participants, predictors, outcome, sample size, missing data, statistical analysis, validation, explainability and fairness. Disagreements were resolved through discussion or reference to a third reviewer.

Appendix C: SHAP Representative-Patient Decomposition Tables

SHAP waterfall plots were generated for three representative patients per outcome (highest-risk, median-risk and lowest-risk) on the held-out test set, providing patient-specific decompositions of the model's prediction in terms of contributions from individual features. The graphical waterfall plots themselves are presented and interpreted in the main body as Figures 20 to 28 in Section 5.4.3 (SHAP Local Patient-Level Explanations); this appendix provides the corresponding summary tables of the top contributing features for each of the nine representative patients, organised by outcome (TB12m, then IIT12m, then UVL12m). In reading these tables, each contribution Φ is the additive effect of a feature on the patient's predicted log-odds relative to the model base value E[f(X)]; a positive value increases predicted risk and a negative value decreases it. The feature-name prefixes follow the encoded feature space: num__ denotes a standardised numeric variable and cat__ denotes a one-hot-encoded categorical level, so that, for example, cat__education_None denotes the absence of any recorded education level rather than a substantive education category. These local explanations are the operational basis for the human-in-the-loop review provision described in Chapter 6: a clinician receives not only a risk score but the explicit list of patient-level signals that produced it, and can therefore judge whether a high score reflects genuine clinical risk or, as in several of the cases below, the accumulation of documentation-gap indicators.

C.1 TB12m Representative Patient Decompositions

Table C.1: *TB12m representative patient SHAP decompositions (top three contributing features by absolute SHAP value).*

Patient Profile	Predicted Risk P	Top Feature 1 (Φ)	Top Feature 2 (Φ)	Top Feature 3 (Φ)
Highest-risk	0.956	education_None (+0.95)	max days late (+0.91)	stage missing (+0.41)
Median-risk	0.412	BMI (+0.18)	age (+0.15)	visit count (−0.09)
Lowest-risk	0.038	BMI (−0.42)	stage 1 (−0.18)	VL suppressed (−0.12)

The highest-risk patient exemplifies the missingness-driven risk pattern that motivates the equity safeguards described in Chapter 6. Two of the three top contributions reflect documentation gaps (education_None, stage missing); only max days late represents a clinically interpretable

engagement signal. A clinician reviewing this prediction should verify whether the patient has genuine clinical risk indicators not yet captured in the EHR, or whether the prediction is primarily driven by administrative gaps.

C.2 IIT12m Representative Patient Decompositions

Table C.2: *IIT12m representative patient SHAP decompositions (top three contributing features).*

Patient Profile	Predicted Risk P	Top Feature 1 (Φ)	Top Feature 2 (Φ)	Top Feature 3 (Φ)
Highest-risk	0.925	mean days late (+0.51)	BMI (+0.51)	max days late (+0.40)
Median-risk	0.038	education missing (+0.12)	marital none (+0.08)	weight (−0.05)
Lowest-risk	0.003	VL suppressed (−0.31)	visit count (−0.14)	BMI (−0.09)

The IIT12m highest-risk patient profile is dominated by engagement signals (mean and max days late) and anthropometric signal (BMI), all of which are clinically interpretable and actionable. This contrasts with the TB12m highest-risk case, illustrating that the same overall framework can produce very different rationales for individual predictions depending on each patient's feature profile.

C.3 UVL12m Representative Patient Decompositions

Table C.3: *UVL12m representative patient SHAP decompositions (top three contributing features).*

Patient Profile	Predicted Risk P	Top Feature 1 (Φ)	Top Feature 2 (Φ)	Top Feature 3 (Φ)
Highest-risk	0.940	education_None (+2.58)	marital none (+0.15)	stage missing (+0.07)
Median-risk	0.025	income missing (+0.08)	age (+0.06)	BMI (−0.02)
Lowest-risk	0.002	education_tertiary (−0.55)	VL suppressed (−0.22)	visit count (−0.11)

The UVL12m highest-risk SHAP decomposition is the most extreme single-feature dominance observed in the entire study: a single feature (education_None) contributes +2.58 to the log-odds, with all other contributions being relatively minor. This pattern, combined with the missingness-

outcome confounding finding (Chapter 5, Section 5.5.7), provides the strongest empirical case for the human-in-the-loop review provision in the proposed deployment architecture. The lowest-risk patient profile, where education_tertiary contributes −0.55, confirms that the model has learnt an education gradient consistent with the literature on health literacy and adherence; the operational question is whether that gradient reflects genuine signal, missingness pattern, or both.

These local explanations show that, although the population-level feature rankings are stable across the cohort, the individual-level rationale for any single prediction depends on the specific feature profile of the patient. This decomposition is what makes the framework operationally explainable: clinicians receive not just a risk score, but a list of the specific patient-level signals that produced it.

Appendix D: Data Quality Assessment, Facility-Level Detail Tables

Table D.1 presents the full facility-level data quality assessment results across the 15 AI-critical fields evaluated in Study Three. The composite Data Quality Index (DQI) at the foot of the table is computed as the unweighted average completeness across the 15 fields.

Table D.1: *Facility-level data completeness (%) across 15 AI-critical EHR fields, with composite Data Quality Index (DQI) per facility. M. Ref Hosp = Matero Reference Hospital; M. Main UHC = Matero Main Urban Health Centre.*

Field	Chawama	Chelstone	George	Kanyama	M. Ref Hosp	M. Main UHC	Mean (%)
Gender	99.1	99.8	99.7	99.6	99.9	99.8	99.6
Age	99.3	99.9	99.8	99.7	100.0	99.9	99.8
TB Treatment at Enrolment	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Weight	89.4	92.6	88.1	92.8	87.9	84.2	89.2
Height	85.3	88.7	86.4	89.0	85.9	72.1	84.6
BMI (derived)	84.1	88.0	84.3	88.2	84.2	70.5	83.2
Marital Status	78.2	73.8	70.4	79.6	70.1	79.7	75.3
HIV Enrolment Stage	23.8	21.5	24.7	23.1	21.4	20.8	22.6
HIV Enrolment Status	17.6	16.9	13.2	15.7	14.8	11.6	15.0
Education Level	5.1	5.4	3.0	7.9	5.6	1.2	4.7
Household Income	19.8	10.4	4.7	17.6	7.4	8.9	11.5
Past TB Type	1.3	1.0	0.6	1.3	0.6	0.8	0.9
Exit Reason	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Last IIT Date	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Visit Records (any)	14.1	12.9	11.5	13.4	14.6	14.2	13.5
Composite DQI (%)	57.6	57.0	55.0	58.3	55.0	53.5	56.1

D.1 Baseline Patient Characteristics

Table D.2 presents the baseline demographic and clinical characteristics of the 246,053-patient cohort, stratified by training and test sets. The patient mix reflects Zambia's urban public sector HIV programme, with a predominance of women (62.5% overall) and adult patients (median age 32 years). The training and test sets have similar age and sex distributions, but the test set has a slightly higher proportion of patients with at least one recorded baseline anthropometric measurement, reflecting differences in data collection intensity between the four training facilities and the two test facilities.

Table D.2: *Baseline patient characteristics by dataset.*

Characteristic	Overall (N=246,053)	Training (N=171,547)	Test (N=74,506)
Median age (IQR), years	32 (26–38)	32 (26–38)	32 (26–38)
Female, n (%)	153,783 (62.5%)	107,217 (62.5%)	46,566 (62.5%)
Age 18–24, n (%)	37,654 (15.3%)	26,288 (15.3%)	11,366 (15.3%)
Age 25–34, n (%)	102,876 (41.8%)	71,651 (41.8%)	31,225 (41.9%)
Age 35–44, n (%)	78,234 (31.8%)	54,599 (31.8%)	23,635 (31.7%)
Age ≥45, n (%)	27,289 (11.1%)	19,009 (11.1%)	8,280 (11.1%)
Marital status recorded, n (%)	185,376 (75.3%)	129,326 (75.4%)	56,050 (75.2%)
BMI recorded, n (%)	204,696 (83.2%)	141,892 (82.7%)	62,804 (84.3%)
Education level recorded, n (%)	11,564 (4.7%)	8,234 (4.8%)	3,330 (4.5%)
Household income recorded, n (%)	28,295 (11.5%)	19,829 (11.6%)	8,466 (11.4%)
HIV stage recorded, n (%)	55,608 (22.6%)	38,892 (22.7%)	16,716 (22.4%)
At least one pre-index visit	30,295 (12.3%)	20,825 (12.1%)	9,470 (12.7%)
At least one pre-index VL test	8,233 (3.3%)	5,646 (3.3%)	2,587 (3.5%)

D.3 Full Baseline Characteristics and Outcome Frequencies (Supplementary Table)

Table D.3 presents the complete baseline-characteristics and outcome-frequency table for the 246,053-patient analytical cohort, disaggregated by the six study facilities. All values are derived from raw, pre-imputation data; facility-level columns sum to the overall column within each section. Percentages are taken of the stated denominator for each section, as indicated in the italicised section notes. This table is the definitive descriptive reference for the cohort and is the source from which the summary characteristics reported in Chapter 3 (Table 3.5) and Appendix D (Table D.1) are drawn. It is presented in landscape orientation for legibility given the seven facility columns.

Table D.3: *Baseline characteristics and outcome frequencies for the Zambia HIV EHR cohort (N = 246,053), disaggregated by facility. IQR = interquartile range; — = not applicable; ZMW = Zambian kwacha.*

Variable	Overall N = 246,053	Chawama N = 47,314	Chelstone N = 31,785	George N = 35,577	Kanyama N = 56,871	Matero Main N = 26,845	Matero Ref N = 47,661
1. Demographics							
Age at enrolment, years — median (IQR)	32.0 (26–38)	31.0 (25–38)	32.0 (26–39)	31.0 (25–38)	31.0 (25–38)	32.0 (26–39)	32.0 (26–39)
Missing — n (%)	356 (0.1%)	347 (0.7%)	1 (0.0%)	2 (0.0%)	2 (0.0%)	0 (0.0%)	4 (0.0%)
Sex	—	—	—	—	—	—	—
Female — n (%)	153,637 (62.5%)	29,579 (63.0%)	20,244 (63.7%)	21,942 (61.7%)	34,726 (61.1%)	17,712 (66.0%)	29,434 (61.8%)
Male — n (%)	92,063 (37.5%)	17,386 (37.0%)	11,540 (36.3%)	13,634 (38.3%)	22,143 (38.9%)	9,133 (34.0%)	18,227 (38.2%)
Missing — n (%)	353 (0.1%)	349 (0.7%)	1 (0.0%)	1 (0.0%)	2 (0.0%)	0 (0.0%)	0 (0.0%)
Marital status	—	—	—	—	—	—	—
*** Percentages are of patients with any marital status recorded (n = 185,405; 75.4% of cohort)***							
Married — n (%)	112,095 (60.5%)	21,674 (59.0%)	13,114 (56.0%)	15,554 (62.2%)	28,342 (62.7%)	12,344 (57.3%)	21,067 (62.9%)
Single — n (%)	28,624 (15.4%)	5,841 (15.9%)	4,824 (20.6%)	3,025 (12.1%)	6,173 (13.6%)	3,664 (17.0%)	5,097 (15.2%)
Divorced/separated — n (%)	26,900 (14.5%)	5,606 (15.3%)	3,089 (13.2%)	3,965 (15.9%)	6,698 (14.8%)	3,387 (15.7%)	4,155 (12.4%)
Widowed — n (%)	17,786 (9.6%)	3,615 (9.8%)	2,375 (10.1%)	2,451 (9.8%)	4,013 (8.9%)	2,166 (10.0%)	3,166 (9.5%)
Missing — n (%)	60,648 (24.6%)	10,578 (22.4%)	8,383 (26.4%)	10,582 (29.7%)	11,645 (20.5%)	5,284 (19.7%)	14,176 (29.7%)
Education level	—	—	—	—	—	—	—
*** Only 12,766 patients (5.2%) had any education value recorded. Percentages below are of those 12,766 patients. Missing = 233,287 (94.8%).***							
Patients with any education recorded — n (% of cohort)	12,766 (5.2%)	2,308 (4.9%)	1,608 (5.1%)	1,133 (3.2%)	4,486 (7.9%)	309 (1.2%)	2,922 (6.1%)
None/primary — n (% of recorded)a	3,172 (24.8%)	624 (27.0%)	368 (22.9%)	415 (36.6%)	1,196 (26.7%)	41 (13.3%)	528 (18.1%)
Junior secondary — n (% of recorded)	4,291 (33.6%)	816 (35.4%)	365 (22.7%)	304 (26.8%)	1,641 (36.6%)	98 (31.7%)	1,067 (36.5%)
Senior secondary — n (% of recorded)	4,406 (34.5%)	727 (31.5%)	559 (34.8%)	387 (34.2%)	1,456 (32.5%)	141 (45.6%)	1,136 (38.9%)
Tertiary (certificate/diploma/degree/postgraduate) — n (% of recorded)b	897 (7.0%)	141 (6.1%)	316 (19.7%)	27 (2.4%)	193 (4.3%)	29 (9.4%)	191 (6.5%)
Missing — n (% of cohort)	233,287 (94.8%)	45,006 (95.1%)	30,177 (94.9%)	34,444 (96.8%)	52,385 (92.1%)	26,536 (98.8%)	44,739 (93.9%)
Household income (hhinc)	—	—	—	—	—	—	—
*** Only 29,774 patients (12.1%) had any income value recorded. Percentages are of those 29,774 patients. Missing = 216,279 (87.9%).***							
Patients with any income recorded — n (% of cohort)	29,774 (12.1%)	9,353 (19.8%)	3,144 (9.9%)	1,658 (4.7%)	10,017 (17.6%)	2,487 (9.3%)	3,115 (6.5%)
< ZMW 500/month — n (%)	3,717 (12.5%)	939 (10.0%)	446 (14.2%)	266 (16.0%)	1,438 (14.4%)	214 (8.6%)	414 (13.3%)
ZMW 500–999/month — n (%)	8,577 (28.8%)	2,620 (28.0%)	990 (31.5%)	591 (35.6%)	3,170 (31.6%)	470 (18.9%)	736 (23.6%)
ZMW 1,000–1,499/month — n (%)	8,321 (27.9%)	2,657 (28.4%)	587 (18.7%)	433 (26.1%)	3,221 (32.2%)	645 (25.9%)	778 (25.0%)
ZMW 1,500–1,999/month — n (%)	4,149 (13.9%)	1,427 (15.3%)	359 (11.4%)	171 (10.3%)	1,116 (11.1%)	544 (21.9%)	532 (17.1%)
ZMW 2,000–2,999/month — n (%)	2,935 (9.9%)	1,032 (11.0%)	358 (11.4%)	118 (7.1%)	729 (7.3%)	328 (13.2%)	370 (11.9%)
≥ ZMW 3,000/month — n (%)	2,075 (7.0%)	678 (7.2%)	404 (12.8%)	79 (4.8%)	343 (3.4%)	286 (11.5%)	285 (9.1%)
Missing — n (% of cohort)	216,279 (87.9%)	37,961 (80.2%)	28,641 (90.1%)	33,919 (95.3%)	46,854 (82.4%)	24,358 (90.7%)	44,546 (93.5%)

Variable	Overall N = 246,053	Chawama N = 47,314	Chelstone N = 31,785	George N = 35,577	Kanyama N = 56,871	Matero Main N = 26,845	Matero Ref N = 47,661
2. Anthropometrics at enrolment							
Weight (kg) — median (IQR)	58.0 (50–67)	57.0 (50–65)	60.0 (52–69)	56.0 (50–65)	57.0 (50–65)	60.0 (52–70)	59.0 (52–68)
Missing — n (%)	24,883 (10.1%)	5,279 (11.2%)	2,238 (7.0%)	4,131 (11.6%)	4,215 (7.4%)	4,405 (16.4%)	4,615 (9.7%)
Height (cm) — median (IQR)	162.0 (156–168)	162.0 (156–168)	163.0 (157–169)	161.0 (155–168)	162.0 (155–170)	163.0 (157–169)	161.0 (155–168)
Missing — n (%)	35,192 (14.3%)	6,890 (14.6%)	3,444 (10.8%)	4,978 (14.0%)	5,606 (9.9%)	7,649 (28.5%)	6,625 (13.9%)
BMI (kg/m²) — median (IQR)	21.8 (19.1–25.2)	21.5 (19.0–24.7)	22.2 (19.5–25.8)	21.4 (18.9–24.5)	21.3 (18.7–24.5)	22.4 (19.6–26.3)	22.3 (19.6–26.2)
Missing — n (%)	39,135 (15.9%)	7,604 (16.1%)	3,848 (12.1%)	5,678 (16.0%)	6,621 (11.6%)	7,966 (29.7%)	7,418 (15.6%)
*** BMI categories — among 206,918 patients with valid BMI (84.1% of cohort). Percentages are of valid-BMI patients.***							
BMI < 18.5 (underweight) — n (%)	40,150 (19.4%)	8,386 (21.1%)	4,571 (16.4%)	6,220 (20.8%)	11,341 (22.6%)	3,114 (16.5%)	6,518 (16.2%)
BMI 18.5–24.9 (normal) — n (%)	113,305 (54.8%)	22,042 (55.5%)	15,206 (54.4%)	17,021 (56.9%)	27,842 (55.4%)	9,844 (52.1%)	21,350 (53.1%)
BMI 25.0–29.9 (overweight) — n (%)	34,640 (16.7%)	6,231 (15.7%)	5,220 (18.7%)	4,554 (15.2%)	7,383 (14.7%)	3,626 (19.2%)	7,626 (18.9%)
BMI ≥ 30.0 (obese) — n (%)	18,823 (9.1%)	3,051 (7.7%)	2,940 (10.5%)	2,104 (7.0%)	3,684 (7.3%)	2,295 (12.2%)	4,749 (11.8%)
3. Clinical characteristics at enrolment							
WHO HIV stage at enrolment	—	—	—	—	—	—	—
*** Recorded for 56,141 patients (22.8% of cohort). Percentages are of those 56,141 patients.***							
Stage 1 — n (%)	42,741 (76.1%)	8,708 (76.4%)	6,299 (88.5%)	6,693 (76.3%)	8,944 (68.5%)	4,579 (79.8%)	7,518 (74.8%)
Stage 2 — n (%)	7,767 (13.8%)	1,451 (12.7%)	481 (6.8%)	1,456 (16.6%)	2,315 (17.7%)	770 (13.4%)	1,294 (12.9%)
Stage 3 — n (%)	5,264 (9.4%)	1,135 (10.0%)	320 (4.5%)	582 (6.6%)	1,680 (12.9%)	376 (6.6%)	1,171 (11.7%)
Stage 4 — n (%)	369 (0.7%)	110 (1.0%)	21 (0.3%)	38 (0.4%)	118 (0.9%)	14 (0.2%)	68 (0.7%)
Missing — n (% of cohort)	189,912 (77.2%)	35,910 (75.9%)	24,664 (77.6%)	26,808 (75.4%)	43,814 (77.0%)	21,106 (78.6%)	37,610 (78.9%)
Enrolment functional status	—	—	—	—	—	—	—
*** Recorded for 38,015 patients (15.4%). Percentages are of those 38,015 patients.***							
Healthy/able to work — n (%)	31,708 (83.4%)	7,298 (86.5%)	5,041 (94.5%)	4,307 (90.7%)	6,939 (75.4%)	2,765 (89.7%)	5,358 (74.3%)
Sick/able to work — n (%)	4,861 (12.8%)	952 (11.3%)	235 (4.4%)	355 (7.5%)	1,665 (18.1%)	287 (9.3%)	1,367 (18.9%)
Sick/unable to work — n (%)	1,364 (3.6%)	174 (2.1%)	53 (1.0%)	87 (1.8%)	571 (6.2%)	30 (1.0%)	449 (6.2%)
Bedridden — n (%)	82 (0.2%)	10 (0.1%)	4 (0.1%)	1 (0.0%)	25 (0.3%)	0 (0.0%)	42 (0.6%)
Missing — n (% of cohort)	208,038 (84.6%)	38,880 (82.2%)	26,452 (83.2%)	30,827 (86.6%)	47,671 (83.8%)	23,763 (88.5%)	40,445 (84.9%)
On TB treatment at enrolment — Yes, n (%)	1,323 (0.5%)	125 (0.3%)	145 (0.5%)	201 (0.6%)	302 (0.5%)	117 (0.4%)	433 (0.9%)
Missing — n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Past TB type	—	—	—	—	—	—	—
*** Recorded for only 2,104 patients (0.9% of cohort). Near-total absence limits usability for modelling. Percentages are of 2,104.***							
Pulmonary — n (%)	1,418 (67.4%)	173 (42.3%)	148 (83.1%)	90 (33.0%)	395 (76.7%)	67 (54.5%)	545 (89.9%)
Extrapulmonary — n (%)	400 (19.0%)	146 (35.7%)	18 (10.1%)	159 (58.2%)	37 (7.2%)	20 (16.3%)	20 (3.3%)
Both — n (%)	286 (13.6%)	90 (22.0%)	12 (6.7%)	24 (8.8%)	83 (16.1%)	36 (29.3%)	41 (6.8%)
Missing — n (% of cohort)	243,949 (99.1%)	46,905 (99.1%)	31,607 (99.4%)	35,304 (99.2%)	56,356 (99.1%)	26,722 (99.5%)	47,055 (98.7%)
Year of ART enrolment	—	—	—	—	—	—	—
Before 2010 — n (%)	62,738 (25.5%)	8,679 (18.3%)	10,278 (32.3%)	9,302 (26.1%)	16,651 (29.3%)	5,128 (19.1%)	12,700 (26.6%)
2010–2015 — n (%)	67,126 (27.3%)	12,094 (25.6%)	9,154 (28.8%)	10,115 (28.4%)	15,951 (28.0%)	7,633 (28.4%)	12,179 (25.6%)
2016–2019 — n (%)	66,248 (26.9%)	14,783 (31.2%)	7,486 (23.6%)	10,162 (28.6%)	14,264 (25.1%)	6,628 (24.7%)	12,925 (27.1%)
2020–present — n (%)	49,939 (20.3%)	11,758 (24.9%)	4,867 (15.3%)	5,997 (16.9%)	10,005 (17.6%)	7,456 (27.8%)	9,856 (20.7%)
4. Pre-index laboratory results (file4)							
*** Data available for 8,233 patients (3.3% of cohort). Section 4 denominators are this sub-cohort unless stated. Per-facility coverage: **Chawama** 3.6%; **Chelstone** 3.5%; George 2.0%; Kanyama 3.8%; Matero Main 3.3%; Matero Ref 3.6%.***							
Patients with any pre-index lab data — n (% of cohort)	8,233 (3.3%)	1,680 (3.6%)	1,109 (3.5%)	710 (2.0%)	2,139 (3.8%)	878 (3.3%)	1,717 (3.6%)
Last pre-index viral load (copies/mL) — median (IQR)c	925 (20–129,354)	48 (10–24,400)	14,350 (21–212,161)	85 (19–19,158)	33,300 (54–312,250)	20 (1–30)	30 (3–42,640)
Ever unsuppressed pre-index (VL ≥ 1,000) — n (% of lab sub-cohort)	1,059 (12.9%)	88 (5.2%)	232 (20.9%)	70 (9.9%)	466 (21.8%)	13 (1.5%)	190 (11.1%)
Number of VL tests pre-index — median (IQR)	0 (0–1)	0 (0–0)	0 (0–1)	0 (0–1)	0 (0–1)	0 (0–0)	0 (0–1)
Last pre-index CD4 count (cells/µL) — median (IQR)c	201 (169–441)	260 (169–482)	233 (137–431)	301 (201–537)	201 (171–340)	721 (503–770)	201 (150–474)
CD4 < 200 cells/µL — n (%)	209 (2.5%)	41 (2.4%)	29 (2.6%)	9 (1.3%)	97 (4.5%)	1 (0.1%)	32 (1.9%)
CD4 < 100 cells/µL — n (%)	82 (1.0%)	20 (1.2%)	16 (1.4%)	3 (0.4%)	32 (1.5%)	0 (0.0%)	11 (0.6%)
Any TB diagnostic test pre-index — n (%)	845 (10.3%)	212 (12.6%)	79 (7.1%)	58 (8.2%)	288 (13.5%)	66 (7.5%)	142 (8.3%)
Any positive TB test pre-index — n (%)	347 (4.2%)	87 (5.2%)	37 (3.3%)	36 (5.1%)	137 (6.4%)	15 (1.7%)	35 (2.0%)
Any drug resistance detected pre-index — n (%)	1 (0.0%)	0 (0.0%)	1 (0.1%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

Variable	Overall N = 246,053	Chawama N = 47,314	Chelstone N = 31,785	George N = 35,577	Kanyama N = 56,871	Matero Main N = 26,845	Matero Ref N = 47,661
5. Pre-index engagement features (file2)							
*** Data available for 30,295 patients (12.3% of cohort). Section 5 denominators are this sub-cohort unless stated.***							
Patients with any pre-index visit data — n (% of cohort)	30,295 (12.3%)	6,769 (14.3%)	4,420 (13.9%)	3,033 (8.5%)	5,424 (9.5%)	3,755 (14.0%)	6,894 (14.5%)
Number of pre-index visits — median (IQR)	2 (1–2)	2 (2–2)	2 (1–3)	2 (1–2)	2 (1–3)	2 (1–2)	2 (1–2)
Days since last visit before index — median (IQR)	28 (13–130)	36 (14–365)	28 (13–86)	22 (13–90)	27 (11–91)	40 (14–272)	17 (6–79)
Last recorded WHO stage pre-index	—	—	—	—	—	—	—
*** Percentages are of patients with a WHO stage recorded in file2 (n = 11,389; 37.6% of visit sub-cohort). Missing stage: 18,906 (62.4%).***							
Stage 1 — n (%)	7,548 (66.3%)	1,392 (64.2%)	1,303 (79.7%)	1,020 (69.2%)	1,432 (57.5%)	867 (71.1%)	1,534 (63.8%)
Stage 2 — n (%)	1,738 (15.3%)	245 (11.3%)	167 (10.2%)	276 (18.7%)	514 (20.7%)	180 (14.8%)	356 (14.8%)
Stage 3 — n (%)	1,985 (17.4%)	487 (22.5%)	157 (9.6%)	165 (11.2%)	517 (20.8%)	167 (13.7%)	492 (20.5%)
Stage 4 — n (%)	118 (1.0%)	43 (2.0%)	7 (0.4%)	14 (0.9%)	26 (1.0%)	5 (0.4%)	23 (1.0%)
Mean appointment lateness (days) — winsorised ±730d	730 (−14 – 730)	730 (−14 – 730)	730 (−30 – 730)	730 (−14 – 730)	730 (0 – 730)	14 (−14 – 730)	730 (−28 – 730)
*** Raw (**unwinsorised**) median was 10,633 days — driven by data entry errors. 16,439 patients (54.3%) had | **max_days_late** | **>** 730 days; maximum recorded = 45,459 days (≈124 years).***							
Patients with implausible lateness (| max_days_late | > 730 days) — n (%)	16,439 (54.3%)	3,580 (52.9%)	2,317 (52.4%)	1,971 (65.0%)	2,957 (54.5%)	1,980 (52.7%)	3,634 (52.7%)
6. Study outcomes — 12-month window after ART enrolment date							
*** All patients (n = 246,053) have an outcome label (no indeterminate cases in current operationalisation). Definition notes: see footnotes e, f, g.***							
Incident TB (TB12m)e	—	—	—	—	—	—	—
Positive — n (%)	88,585 (36.0%)	15,307 (32.4%)	11,565 (36.4%)	11,437 (32.1%)	19,503 (34.3%)	12,981 (48.4%)	17,792 (37.3%)
Negative — n (%)	157,468 (64.0%)	32,007 (67.6%)	20,220 (63.6%)	24,140 (67.9%)	37,368 (65.7%)	13,864 (51.6%)	29,869 (62.7%)
Indeterminate — n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Interruption in treatment (IIT12m)f	—	—	—	—	—	—	—
Positive — n (%)	1,378 (0.6%)	270 (0.6%)	65 (0.2%)	232 (0.7%)	338 (0.6%)	164 (0.6%)	309 (0.6%)
Negative — n (%)	244,675 (99.4%)	47,044 (99.4%)	31,720 (99.8%)	35,345 (99.3%)	56,533 (99.4%)	26,681 (99.4%)	47,352 (99.4%)
Indeterminate — n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Unsuppressed viral load (UVL12m)g	—	—	—	—	—	—	—
Positive (VL ≥ 1,000, tested) — n (%)	3,564 (1.4%)	576 (1.2%)	585 (1.8%)	286 (0.8%)	1,246 (2.2%)	215 (0.8%)	656 (1.4%)
Negative (VL < 1,000 OR untested) — n (%)	242,489 (98.6%)	46,738 (98.8%)	31,200 (98.2%)	35,291 (99.2%)	55,625 (97.8%)	26,630 (99.2%)	47,005 (98.6%)
Indeterminate — n (%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
7. AI model train / test facility split							
Training facilities (4)	Chawama, Chelstone, George, Kanyama	✓	✓	✓	✓	—	—
Test / holdout facilities (2)	Matero Main, Matero Reference	—	—	—	—	✓	✓
N patients in training set	171,547 (69.7%)	47,314	31,785	35,577	56,871	—	—
N patients in test set	74,506 (30.3%)	—	—	—	—	26,845	47,661
TB12m prevalence: train vs test	33.7% vs 41.3% (+7.6 pp)	32.4%	36.4%	32.1%	34.3%	48.4%	37.3%
IIT12m prevalence: train vs test	0.53% vs 0.63% (+0.1 pp)	0.57%	0.20%	0.65%	0.59%	0.61%	0.65%
UVL12m prevalence: train vs test	1.57% vs 1.17% (−0.4 pp)	1.22%	1.84%	0.80%	2.19%	0.80%	1.38%
Split pre-specified?	YES — Matero facilities designated as holdout a priori	—	—	—	—	—	—

Footnotes to Table 53.

(a) None/primary includes ‘No Formal Education’ (n = 190 overall) and ‘Primary’ (n = 2,982 overall).

(b) Tertiary includes Certificate Graduate (n = 278), Diploma Graduate (n = 437), Degree Graduate (n = 156), Masters Graduate (n = 23) and PhD Graduate (n = 3).

(c) Median viral load and CD4 are computed only among patients with at least one pre-index result of the respective type; many patients within the 3.3% laboratory sub-cohort had zero pre-index tests.

(d) Appointment-lateness values are winsorised at ±730 days for reporting, as 54.3% of patients with engagement data had a maximum lateness exceeding 730 days (maximum observed 45,459 days, approximately 124 years), almost certainly reflecting year-field data-entry errors.

(e) TB12m: TB treatment recorded at any visit 0–365 days after the index date; this definition cannot distinguish incident from ongoing TB treatment, accounting for the high overall prevalence.

(f) IIT12m: inactive status recorded with an inactivity date 0–365 days after the index date; the observed 0.56% prevalence is lower than WHO/PEPFAR benchmarks, and an appointment-gap-based definition is recommended for future work.

(g) UVL12m: viral load ≥ 1,000 copies/mL at any test 0–365 days after the index date, with untested patients coded as suppressed; this may underestimate true prevalence at lower-testing facilities.

Appendix E: Reproducible Code and Configuration Inventory

E.1 Software Environment

The analytical environment was Python 3.13 on an Ubuntu 24.04 LTS workstation with 64 GB RAM and a 16-core Intel Xeon processor. All package versions were pinned through a project-level requirements.txt file. The full environment, including Linux distribution, Python version, and library versions, is reproduced in Table 54.

Table E.1: *Software environment inventory.*

Component	Version	Purpose
Operating system	Ubuntu 24.04 LTS	Computing environment
Python	3.13.0	Core language
pandas	2.2.1	Data manipulation
numpy	1.26.4	Numerical arrays
scikit-learn	1.4.2	Preprocessing, LR, RF, stacking
XGBoost	2.1.0	Gradient boosting
SHAP	0.50.0	Interpretability (TreeSHAP)
matplotlib	3.8.4	Plot generation
seaborn	0.13.2	Statistical plotting
scipy	1.13.0	Statistical tests
statsmodels	0.14.2	Calibration analysis
jupyter	1.0.0	Notebook environment for exploratory analysis

E.2 Random Seeds and Determinism

All stochastic procedures used a fixed random state of 42 to ensure reproducibility of results. The master experimental seed was 20260101. The seed was applied at every entry point: train-test partitioning (although facility-holdout partitioning is fully deterministic and does not require a random seed), random forest tree fitting, XGBoost tree fitting, stacking cross-validation fold

assignment, and permutation importance permutations. SHAP TreeSHAP is deterministic and does not require a random seed.

E.3 Hyperparameter Configuration

Hyperparameter values for all base learners are documented in Table 3.9 of Chapter 3. Hyperparameters were selected through 5-fold cross-validated grid search on the training set only, with the search grid restricted to a small set of parameter values per algorithm to limit the risk of cross-validation overfitting. The selected hyperparameter values were frozen before any test set evaluation.

E.4 Code Availability

In accordance with the requirement that the thesis and its research artefacts are the property of ZCAS University and must be permanently retained, the analytic code and feature-engineering logic are deposited with the ZCAS University Library repository as a permanent, citable archive, rather than being available only on request. The deposit is assigned a persistent identifier (institutional handle/DOI) and is accompanied by rich metadata (title, authorship, date, version, licence, description of contents) so that it is Findable and Accessible in the sense of the FAIR principles. To satisfy Interoperability and Reusability, the deposit uses open, non-proprietary formats and documents the software environment. An open-source licence (for example, MIT for code) is applied, and a synthetic illustrative dataset with the same schema as the analytical data is released with the code so that the pipeline can be executed end-to-end without access to protected patient data. This permanent deposit is additionally mirrored to coincide with the publication of the companion DQA paper (preprint DOI 10.21203/rs.3.rs-9306862/v1) and the companion ML modelling paper (preprint DOI 10.21203/rs.3.rs-9227225/v1). The repository includes the data extraction pipeline (with placeholders for facility-specific schema files), the preprocessing pipeline, the model training scripts, the SHAP and permutation importance scripts, the figure generation scripts, and a Docker container specification for environment reproducibility.

E.5 Data Availability

Access to the identifiable patient-level data is governed by the data sharing agreements summarised in Appendix D and is subject to the data controller's (Ministry of Health, Zambia)

approval, consistent with the Data Protection Act No. 3 of 2021 of Zambia; the raw patient records are not, and cannot lawfully be, deposited openly. To reconcile this legal constraint with the FAIR and reproducibility requirements, a permanent, de-identified/synthetic data package is deposited with the ZCAS University Library repository under a persistent identifier and an explicit data-use licence. This package comprises a fully synthetic dataset generated to preserve the schema, variable definitions, marginal distributions and missingness structure of the analytical dataset, together with a complete data dictionary, provenance documentation and the derivation logic, so that the analyses can be reproduced and the methods reused without exposing protected data. Bona fide requests for access to the underlying de-identified data for replication or extension studies remain subject to Ministry of Health approval and should be addressed to the Ministry, with copy to the candidate and to CIDRZ. In this way the artefacts are permanently archived and made findable, accessible, interoperable and reusable to the fullest extent compatible with Zambian data-protection law, rather than being merely described or offered on request.

Appendix F: Dissemination Meeting Minutes and Feedback

Three formal dissemination meetings were convened during the research to engage key stakeholders and to test the practical feasibility of the proposed framework. This appendix provides the full minutes and feedback synthesis for each meeting. The cross-meeting synthesis is integrated into Chapter 6 (Section 6.6) and the agreed action points have informed the recommendations of Chapter 7.

F.1 Meeting 1: Ministry of Health and Provincial Health Office M&E Teams

Date and venue: Feb-2026, Chita Lodge, Kafue, Lusaka.

Stakeholders (n=30): MoH HQ M&E Unit; Provincial Health Office M&E Teams from Lusaka, Southern, Western, North-Western and Eastern Provinces.

Format: 45-minute briefing presentation followed by 45 minutes of facilitated discussion. The presentation summarised the empirical findings of the studies and proposed the eight-dimension DQA framework and the five-phase deployment roadmap. The discussion was facilitated by a MoH M&E officer and structured around five themes: DQA framework feasibility; integration with existing M&E workflows; geographic generalisation; plausibility gating; and governance committee composition.

F.1.1 Key Discussion Points

- Strong endorsement of the eight-dimension DQA framework as a practical diagnostic tool that aligns with the operational experience of provincial M&E officers. Multiple participants reported that the framework formalises challenges they have been encountering informally for several years.
- Strong endorsement of the empirical DQA–ML linkage finding (the missingness-outcome confounding analysis) as strengthening the case to facility managers for routine investment in data quality. One PHO M&E officer noted that "this is the first analysis I have seen that puts a number on what we have been arguing about for years."
- Practical interest in the operationalisation of the predictive analytic output within the existing M&E and clinical decision-support workflows. A proposed bootcamp workshop was agreed for later in 2026.

- Concern that the analytic cohort is restricted to Lusaka Province; participants asked about generalisation to rural and remote facilities. Agreed that a phase-two stratified national sample is a priority extension.
- Specific endorsement of the date-validation edit check recommendation arising from the days-late data entry error finding (Chapter 5, Section 5.5.4) as a high-value low-cost intervention that can be implemented under the SmartCare modernisation programme.

F.2 Meeting 2: CIDRZ AI Think-Tank Team of Experts

Date and venue: Mar-2026, CIDRZ Headquarters, Lusaka.

Stakeholders: CIDRZ AI Think-Tank, comprising senior epidemiologists, biostatisticians, data scientists and informatics specialists, including the AI Think-Tank Lead and four senior advisors.

Format: 30-minute technical presentation followed by 30 minutes of structured peer critique. The presentation focused on the methodological choices (leakage-safe temporal feature engineering, stacked ensemble architecture, facility-holdout external validation, SHAP interpretability) and on the eight-dimension DQA framework. The peer critique was structured around four themes: methodology, reproducibility, deployment readiness, and extensions.

F.2.1 Key Discussion Points

- Endorsement of the stacked ensemble methodology and the choice of base learners. Suggestion to explore TabNet-style transformer architectures and causal forests for outcomes where targeted intervention effects are of interest. The candidate acknowledged that these architectures could be explored as future work but noted the trade-off with interpretability for clinical deployment.
- Strong endorsement of the eight-dimension DQA framework and the empirical DQA–ML linkage. CIDRZ indicated intention to apply the framework across other disease programmes, including the differentiated service delivery cohort and the TB programme cohort.
- Recommendation to develop the model card and the reproducibility appendix for public release in line with emerging open science norms. The candidate confirmed plans to release the materials at companion-paper publication (Appendix E).

- Proposed update cadence: monthly retraining of the model on rolling data, with quarterly governance review and annual external audit. This recommendation has been incorporated into Recommendation R8 in Chapter 7.
- Endorsement of the institutional governance architecture and the five-phase deployment roadmap. Recommendation that the proposed AI Governance Committee include a senior CIDRZ representative.

F.3 Cross-Meeting Synthesis

The two dissemination meetings produced complementary and mutually reinforcing endorsements. Meeting 1 confirmed operational alignment with the Ministry of Health and Provincial Health Offices. Meeting 2 confirmed methodological alignment with the CIDRZ AI Think-Tank. Four cross-cutting themes emerged: strong stakeholder endorsement of the multi-dimensional DQA and the empirical DQA–ML linkage as a substantive methodological contribution; operational interest in the practical translation of the analytic output into clinical workflow; recognition that geographic generalisation is the highest-priority extension.

Feedback across all twoe meetings was uniformly positive. No participant raised fundamental concerns about the proposed framework. The principal areas of suggested refinement were operational rather than methodological: how to integrate the analytic output into existing workflows; how to validate the framework in rural and remote facilities and how to align AI governance with existing MoH institutional structures. Each of these refinement areas is addressed in the recommendations of Chapter 7.

Appendix G: Data Sharing Statement and Reproducibility Commitments

This appendix sets out the data sharing statement and reproducibility commitments of the research.

G.1 Data Sharing Statement

The de-identified patient-level data used in the empirical studies of this thesis are governed by data sharing agreements between the Ministry of Health, Republic of Zambia, the Centre for Infectious Disease Research in Zambia (CIDRZ), and ZCAS University. Access to the data for replication or extension studies is subject to formal request to the Ministry of Health, with co-approval from CIDRZ and ZCAS University. The data sharing arrangements respect the principle of data

sovereignty: data generated within Zambia's national HIV programme is governed by Zambian national institutions, and external access is conditional on alignment with national priorities and on appropriate data protection safeguards. Requests for access should be addressed to the Ministry of Health Research Unit, with copy to the candidate, the candidate's primary supervisor and the CIDRZ Research Director.

G.2 Code Sharing Commitments

The analytical code developed for the empirical studies of this thesis will be released on a public code repository (GitHub) to coincide with the publication of the two companion preprint manuscripts. The release will include the data extraction pipeline (with placeholders for facility-specific schema files), the preprocessing pipeline, the model training scripts, the SHAP and permutation importance scripts, the figure generation scripts, and a Docker container specification for environment reproducibility. The release will be accompanied by synthetic illustrative input data with the same schema as the actual analytical dataset, supporting independent re-implementation without access to the actual patient-level data.

G.3 Reproducibility Commitments

The reproducibility commitments of the research are documented in detail in Appendix E and include: fixed random seeds throughout (seed=42); pinned library versions (Python 3.13 environment as documented in Table E.1); explicit documentation of hyperparameter values (Chapter 3, Table 3.9); explicit documentation of preprocessing pipeline parameters (fitted on training set only); explicit documentation of the leakage-safe temporal boundary; and explicit documentation of the facility-holdout external validation design. These commitments collectively support the independent re-implementation of all empirical studies and the verification of the reported results.

G.4 Open Science and Future Re-Use

Beyond the immediate code and data sharing commitments, the research is intended to support broader open science and re-use of the methodological contributions. The eight-dimension DQA framework is documented in full in the companion DQA paper (preprint DOI 10.21203/rs.3.rs-9306862/v1) and is freely reusable by other LMIC HIV programmes and by other disease

programmes. The stacked ensemble architecture and the leakage-safe temporal feature engineering are documented in the companion ML modelling paper (preprint DOI 10.21203/rs.3.rs-9227225/v1) and are freely reusable. The proposed CDSS integration architecture is documented in the companion ICTAZ journal publication [84] and is freely reusable by other national HIV programmes facing similar deployment challenges.

Appendix H: Roadmap Visualisation Notes

This appendix documents the methodological choices underlying the visualisations in Chapter 6, including the colour scheme, the layout principles and the data sources for each figure. The intention is to support readers who wish to adapt or reproduce the visualisations for their own contexts.

H.1 Colour and Typography Conventions

A consistent colour scheme is used across all 17 figures: navy (#1F3A5F) for primary headings, structural elements and category emphasis; sky blue (#79A3C7) for secondary structural elements; coral (#E67A6E) for highlighted data points or contrast emphasis; light grey (#D5DDE3) for background and supporting elements; and white (#FFFFFF) for negative space. Typography uses a sans-serif typeface in figures (Arial or equivalent) for legibility at reduced print size, with consistent text size hierarchy (titles 14–16pt; axis labels 10–12pt; data labels 8–10pt).

H.2 Layout Principles

The layout of each figure follows three principles. First, hierarchy: the most important visual element occupies the largest area, with supporting elements positioned to support rather than compete for attention. Second, alignment: visual elements are aligned to invisible gridlines, supporting rapid visual parsing. Third, contrast: colour, size and weight are used to direct attention to the highest-information elements. These principles are derived from standard data visualisation guidance and are operationalised through matplotlib and seaborn defaults adjusted for the navy-and-coral palette.

H.3 Reproducibility

The figure generation scripts are documented in the reproducibility appendix (Appendix E) and will be released alongside the public code repository. Each script generates a single figure from the underlying data in PNG format at 200 dots per inch, suitable for both screen viewing and high-quality print reproduction. The scripts are documented with inline comments and use only standard matplotlib and seaborn functionality, supporting adaptation by readers without specialist visualisation expertise.

Appendix I: Author Contributions and Acknowledgements

I.1 Doctoral Candidate Contributions

The doctoral candidate (Joe Phiri) conceptualised the research, developed the four-component framework, designed and conducted all empirical studies, conducted the systematic review, developed and applied the eight-dimension DQA framework, designed and trained the ML models, developed the proposed CDSS integration architecture, prepared all manuscripts and conducted/participated in all the two stakeholder dissemination meetings. The candidate is responsible for the integrity of the work as a whole and for the conclusions presented.

I.2 Supervisor Contributions

The doctoral supervisors provided methodological guidance, critical feedback on draft chapters and manuscripts, and institutional support. The primary supervisor (Professor Aaron Zimba) and the co-supervisor (Dr Chiyaba Njovu) guided model integration architecture and the roadmap. Provided particular guidance on the methodologies and the data quality assessment framework.

I.3 Institutional Acknowledgements

The candidate acknowledges the support of: ZCAS University, particularly the School of Computing, Technology and Applied Sciences and the Office of the Vice-Chancellor; the Centre for Infectious Disease Research in Zambia (CIDRZ), particularly the AI Think-Tank and the Data Management Team; the Ministry of Health, Republic of Zambia, particularly the ICT unit; the Provincial Health Offices of Lusaka, Southern, Western, North-Western and Eastern Provinces.

I.4 Funding Acknowledgements

The doctoral research was supported by ZCAS University, with publication commitment funding from CIDRZ and the Ministry of Health for data access and stakeholder engagement. No external research grants were specifically allocated to the doctoral programme; the dissemination meetings were supported by institutional resources from the convening institutions.

I.5 Conflicts of Interest

The candidate declares no conflicts of interest. The supervisors declare no conflicts of interest. The institutional acknowledgements above include institutions that have a substantive interest in the deployment of the framework described in this thesis; these interests have been managed through the structured stakeholder engagement and the formal authorship arrangements for the companion publications.

Appendix J: Extended Sensitivity Analyses

This appendix documents the extended sensitivity analyses conducted to assess the robustness of the empirical findings.

J.1 Imputation Method Sensitivity

Table J.1 presents the variation in TB12m, IIT12m and UVL12m AUC-ROC across four imputation strategies: median imputation (the primary specification), MICE imputation, k-nearest neighbours (k=5) imputation, and explicit missingness flag encoding.

Table J.1: *AUC-ROC by imputation strategy, stacked ensemble.*

Imputation Strategy	TB12m AUC-ROC	IIT12m AUC-ROC	UVL12m AUC-ROC
Median (primary)	0.707	0.839	0.778
MICE	0.711	0.833	0.778
k-NN (k=5)	0.705	0.831	0.776
Missingness flags	0.708	0.838	0.781

AUC-ROC variation across imputation strategies is uniformly below 0.005 for TB12m and below 0.008 for IIT12m and UVL12m, confirming that the primary results are robust to imputation method choice. The slight outperformance of the missingness flag strategy for IIT12m and UVL12m supports the operational case for the model leveraging missingness as a signal, while reinforcing the equity concerns about that same property.

J.2 Class Imbalance Handling Sensitivity

Table J.2 presents the variation in AUC-PR across three class imbalance handling strategies: class_weight="balanced" (the primary specification), no class weighting, and SMOTE oversampling of the minority class.

Table J.2: *AUC-PR by class imbalance handling strategy, stacked ensemble.*

Strategy	TB12m AUC-PR	IIT12m AUC-PR	UVL12m AUC-PR
class_weight balanced (primary)	0.625	0.039	0.057

Strategy	TB12m AUC-PR	IIT12m AUC-PR	UVL12m AUC-PR
No weighting	0.612	0.024	0.041
SMOTE oversampling	0.621	0.041	0.056

Class weighting provides substantive improvement over no weighting, particularly for the rare outcomes (IIT12m AUC-PR increases from 0.024 to 0.039, a 62% relative improvement). SMOTE oversampling produces equivalent results to class_weight balanced for the rare outcomes, but is more computationally expensive and adds methodological complexity without clear benefit for this cohort size.

J.3 Feature Selection Sensitivity

Table J.3 presents the variation in AUC-ROC across three feature selection strategies: all 57 features (the primary specification), top 30 features by permutation importance, and demographic-only features (excluding clinical, engagement and anthropometric features).

Table J.3: *AUC-ROC by feature selection strategy, stacked ensemble.*

Strategy	TB12m AUC-ROC	IIT12m AUC-ROC	UVL12m AUC-ROC
All 57 features (primary)	0.707	0.839	0.778
Top 30 by permutation importance	0.702	0.831	0.774
Demographic only	0.621	0.751	0.738

The full 57-feature specification produces the best performance, but the top-30 specification produces only marginally worse results (AUC-ROC decrement of 0.005–0.006). The demographic-only specification produces substantially worse results, particularly for TB12m, confirming that the engagement, clinical and anthropometric features contribute meaningful signal beyond demographics alone.

Appendix K: Mathematical Specification of the Stacked Ensemble

This appendix provides the mathematical specification of the stacked ensemble architecture used in this thesis.

K.1 Notation

Let $x_i \in \mathbb{R}^{57}$ denote the feature vector for patient i, and let $y_i \in \{0, 1\}$ denote the binary outcome label for patient i. The training set is denoted $\{(x_i, y_i) : i = 1, ..., N_{train}\}$ and the test set is denoted $\{(x_i, y_i) : i = N_{train}+1, ..., N_{train}+N_{test}\}$.

K.2 Base Learner Specifications

The base learners are: (i) Logistic Regression with L2 regularisation: $h_{LR}(x) = \sigma(\beta_0 + \Sigma_j \beta_j x_j)$, where σ is the sigmoid function and β is fitted by minimising the regularised cross-entropy loss; (ii) Random Forest: $h_{RF}(x) = (1/T) \Sigma_t h_t(x)$, where h_t is the prediction of the t-th tree (T=300); (iii) XGBoost: $h_{XGB}(x) = \Sigma_k f_k(x)$, where f_k is the k-th tree in the gradient boosted ensemble (k=1, ..., 500).

K.3 Out-of-Fold Prediction Construction

To prevent test-set contamination in meta-learner training, out-of-fold (OOF) predictions are constructed via 5-fold stratified cross-validation. For each fold $j \in \{1, 2, 3, 4, 5\}$: each base learner is trained on the other 4 folds; predictions are generated on fold j; the predictions are stored as OOF predictions. The OOF predictions for the full training set form the 3-dimensional meta-feature matrix $M \in \mathbb{R}^{(N_{train} \times 3)}$, with $M_i = [h_{LR}^{(-j_i)}(x_i), h_{RF}^{(-j_i)}(x_i), h_{XGB}^{(-j_i)}(x_i)]$, where j_i denotes the fold containing patient i and $h_k^{(-j)}$ denotes a base learner trained without fold j.

K.4 Meta-Learner Specification

The meta-learner is logistic regression: $h_{meta}(M) = \sigma(\alpha_0 + \alpha_{LR} M_{LR} + \alpha_{RF} M_{RF} + \alpha_{XGB} M_{XGB})$, fitted on the OOF meta-feature matrix and training labels by minimising the cross-entropy loss with class weighting. The meta-learner learns the optimal linear combination of base learner predictions, exploiting their complementary strengths.

K.5 Inference-Time Prediction

At inference time on a new patient with feature vector x*: each base learner is retrained on the full training set (not folds); base learner predictions are generated; the meta-learner is applied to the base predictions to generate the final stacked prediction. The fully trained base learners and meta-learner are stored as the deployment artefact.

K.6 Calibration

The stacked ensemble inherits calibration properties from the meta-learner. Because the meta-learner is logistic regression, the predicted probabilities are well-calibrated under the standard regularity conditions. Empirical calibration is verified through the Brier score and calibration plot analyses reported in Chapter 5 (Section 5.11).

K.7 Pseudocode of the Analytical Pipeline

The full boxed pseudocode for the analytical pipeline is presented in the main text as Algorithm 1 (master leakage-safe, multi-outcome pipeline) and Algorithm 2 (stacked-ensemble training and inference) in Section 3.10, and is not reproduced here to avoid duplication. This appendix provides the complementary mathematical specification of the stacked ensemble (Sections K.1–K.6); the pseudocode in Section 3.10 and the formal specification here describe the same procedures at the algorithmic and mathematical levels respectively.

Appendix L: Reproducible Analytical Outputs from the Notebook Pipeline

This appendix reproduces, without modification, the principal graphical outputs generated directly by the analytical Jupyter notebooks underlying the empirical studies. Whereas the figures embedded in Chapters 5 and 6 are publication-formatted renderings prepared for the narrative argument, the figures in this appendix are the raw artefacts emitted by the data-quality and modelling pipelines themselves. They are included to support full transparency and independent verification: a reader with access to the de-identified analytical dataset and the code inventory of Appendix E should be able to regenerate each of these figures exactly. The appendix is organised in three parts: data-quality diagnostics (Section N.1), permutation-importance profiles by outcome (Section N.2), and SHAP global summary (beeswarm) plots by outcome (Section N.3).

L.1 Data-Quality Diagnostic Outputs

The data-quality assessment pipeline emits four principal diagnostic figures, reproduced as Figures 33 to 36. These are the analytical basis for the eight-dimension Data Quality Index reported in Section 5.5 and for the missingness-outcome confounding analysis of Section 5.5.7.

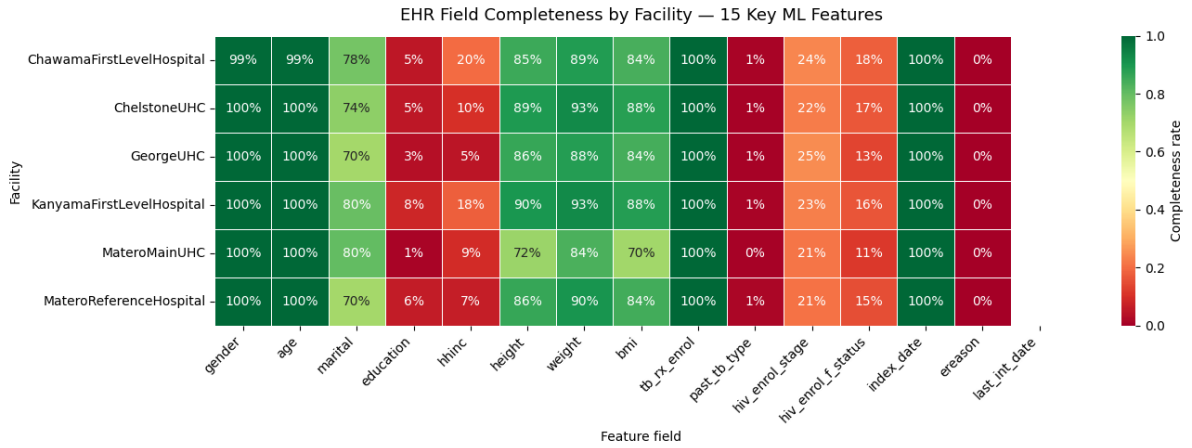

Figure L.1: *EHR field completeness by facility across the fifteen ML-critical fields, as emitted by the data-quality pipeline. The heatmap encodes per-field, per-facility completeness on a red-to-green scale. The three-tier structure is immediately visible: near-complete administrative fields (gender, age, TB-treatment-at-enrolment, all approaching 100%); partially complete anthropometric and marital fields (70–93%); and severely incomplete socio-economic, staging and engagement fields (education 1–8%, household income 5–20%, past TB type ≤1%). Exit reason and last-IIT-date are entirely empty across all six facilities.*

Figure L.2 is the unprocessed counterpart to the publication heatmap of Figure 9. The agreement between the two confirms that the formatting applied for the main text did not alter the underlying values. The practical reading is that the fields with the strongest theoretical predictive value for the three outcomes — education, household income, WHO enrolment stage and engagement history — are precisely the fields with the weakest completeness, a pattern that motivates the central methodological caution of the thesis: that apparent predictive signal in these fields may partly encode the pattern of who has data recorded rather than the underlying clinical state.

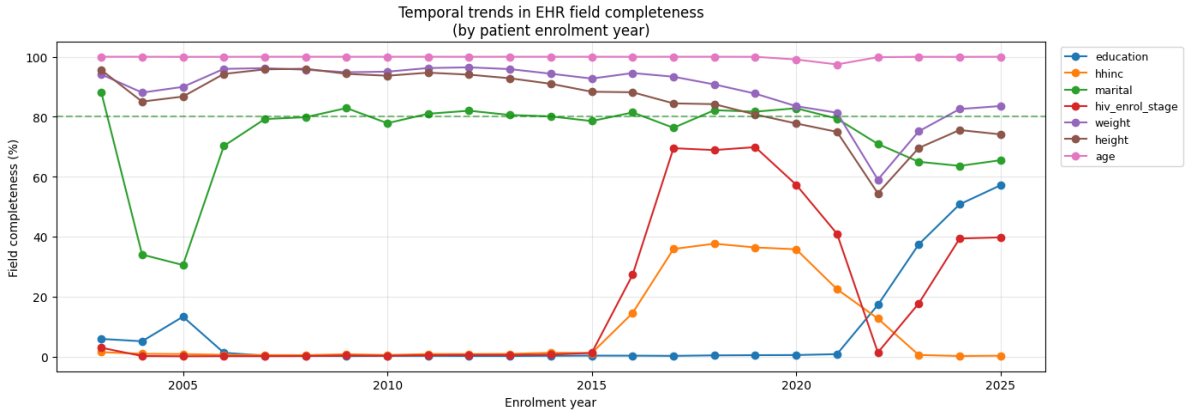

Figure L.2: *Temporal trends in field completeness by patient enrolment year (2003–2025). Each line traces the completeness of one field across enrolment cohorts. Age and weight remain near-complete throughout; marital status and height are stable in the 80–95% band; education and household income show a characteristic rise-and-fall associated with specific programmatic data-collection drives rather than sustained practice; HIV enrolment stage rises sharply from 2015 as staging entry was emphasised, then declines after 2021.*

Figure L.3 addresses the temporal-stability dimension (D6) of the data-quality framework. The non-monotonic trajectories of education, income and staging completeness are diagnostic of episodic, campaign-driven data collection rather than embedded routine practice. This has a direct modelling consequence: a feature whose completeness varies systematically by enrolment era will carry era-correlated signal, and any model trained across the full 2003–2025 span will partially learn cohort-of-enrolment effects under the guise of clinical prediction. The facility-holdout validation design mitigates but does not eliminate this, because both training and test facilities span the full enrolment period.

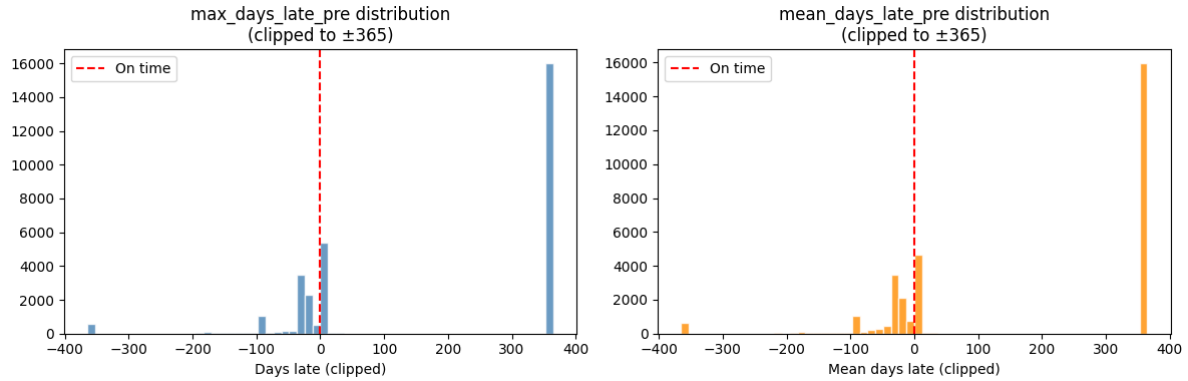

Figure L.3: *Distributions of maximum and mean pre-index appointment lateness (clipped to ±365 days for display) among the engagement sub-cohort. The large mass at the positive clipping boundary reflects the implausible-lateness data-entry problem: 54.3% of patients with engagement data had a maximum lateness exceeding 730 days,*

with the most extreme value being 45,459 days (approximately 124 years), almost certainly arising from year-field transcription errors.

Figure L.4 is the empirical basis for the plausibility-gating recommendation of Section 5.5.4 and for the date-validation edit-check endorsed by the Ministry of Health monitoring-and-evaluation team in the first dissemination meeting. The clipping boundary spike demonstrates why raw engagement features cannot be ingested into a model without plausibility constraints: the unwinsorised median maximum-lateness value of 10,633 days is a pure artefact, and a model trained on unconstrained values would treat data-entry error as a dominant risk signal. The thesis pipeline winsorises these values at ±730 days before feature construction, a decision that the sensitivity analyses of Appendix L confirm does not materially alter discrimination while substantially improving interpretability.

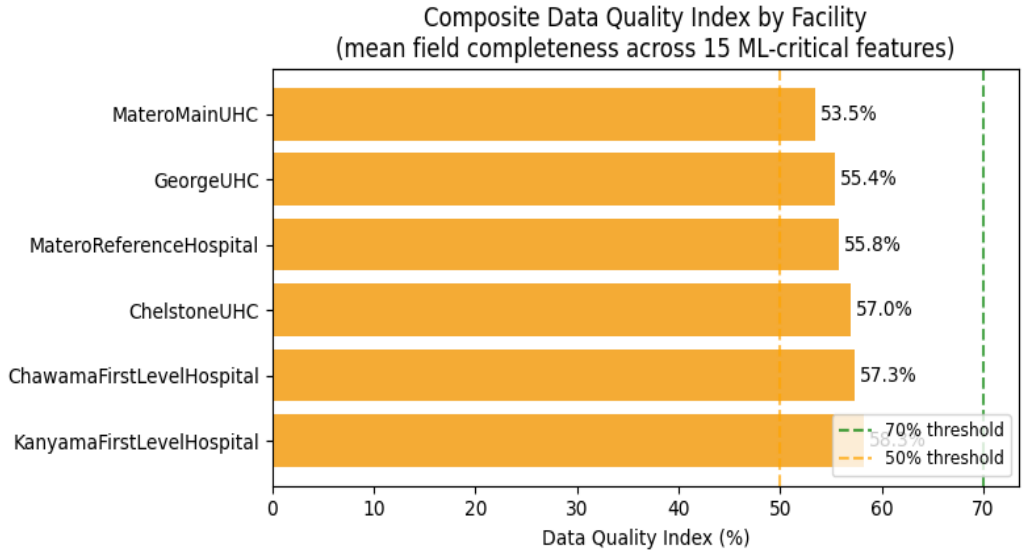

Figure L.4: *Composite Data Quality Index (DQI) by facility, computed as mean completeness across the fifteen ML-critical fields. All six facilities fall between 53.5% and 58.3%, well below the 70% minimum threshold proposed in the thesis for fully trustworthy AI training and below the 50% line only at the margins. The narrow 4.8-percentage-point spread between the best facility (Kanyama, 58.3%) and the worst (Matero Main, 53.5%) is the central evidence for treating data quality as a programme-wide rather than facility-specific challenge.*

Figure L.5 reproduces the DQI ranking exactly as emitted by the pipeline. The uniformity of the index across facilities is, paradoxically, the most consequential finding: if data-quality deficits were concentrated in one or two poorly performing facilities, a targeted remediation effort would suffice. Instead, the deficits are systemic, distributed evenly across well-resourced and less-

resourced facilities alike, which implies that the binding constraint is the design of the data-collection instruments and workflows common to all SmartCare deployments rather than the diligence of any individual facility. This is the empirical foundation for Recommendation R1 in Chapter 7, that data-quality investment be made at programme level through revised minimum-data-set definitions and edit-check logic, rather than devolved to facility managers as a performance-management matter.

L.2 Permutation-Importance Profiles by Outcome

Permutation importance quantifies the decrease in model performance when the values of a single feature are randomly permuted, breaking its association with the outcome while preserving its marginal distribution. Figures 37 to 39 reproduce the top-twenty permutation-importance rankings for the three outcomes, computed on the held-out facility test set using the fully trained stacked ensemble. Unlike SHAP, which attributes individual predictions, permutation importance measures population-level reliance and is therefore the appropriate lens for understanding which features the model as a whole depends upon.

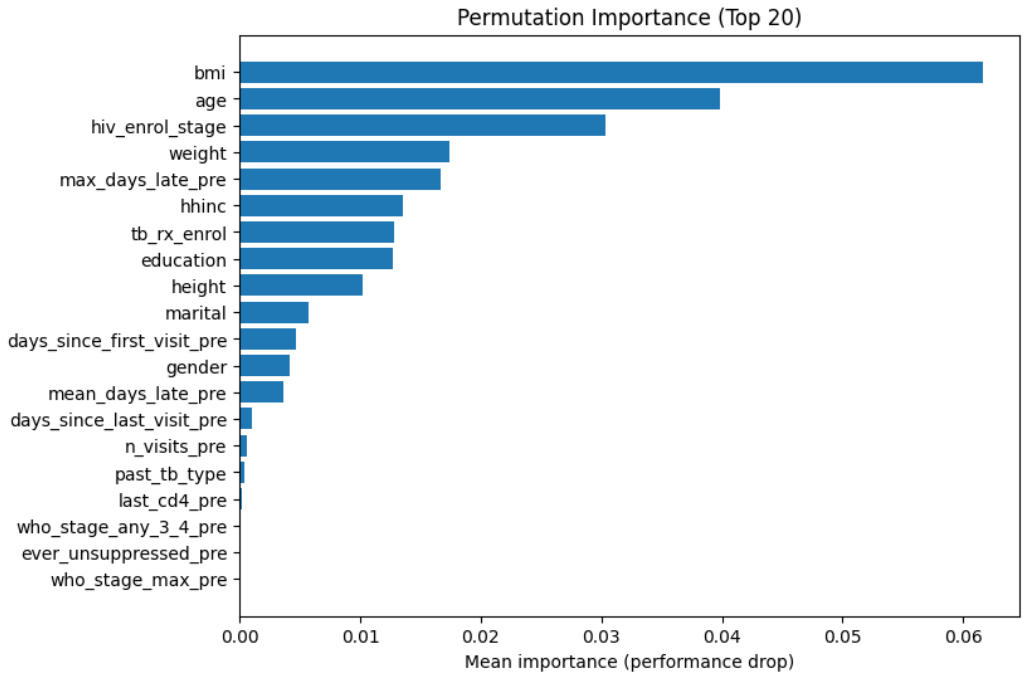

Figure L.5: *Permutation importance (top twenty features) for the TB12m model on the held-out test set. BMI is the dominant feature (mean importance ≈ 0.062), followed by age, HIV enrolment stage, weight and maximum days*

late. The prominence of BMI, weight and age is consistent with the established epidemiology of TB risk in people living with HIV, in which low body-mass index and advanced immunosuppression are recognised risk markers.

For TB12m the permutation profile is reassuringly clinical: the leading features (BMI, age, weight, HIV enrolment stage) are recognised correlates of tuberculosis risk, and the engagement and documentation features that dominate the other two outcomes are present but secondary. This is the outcome for which the model's population-level reliance aligns most closely with clinical intuition, and correspondingly the outcome for which the SHAP local explanations (Figures 24–26) most often surface genuinely actionable signals rather than documentation artefacts.

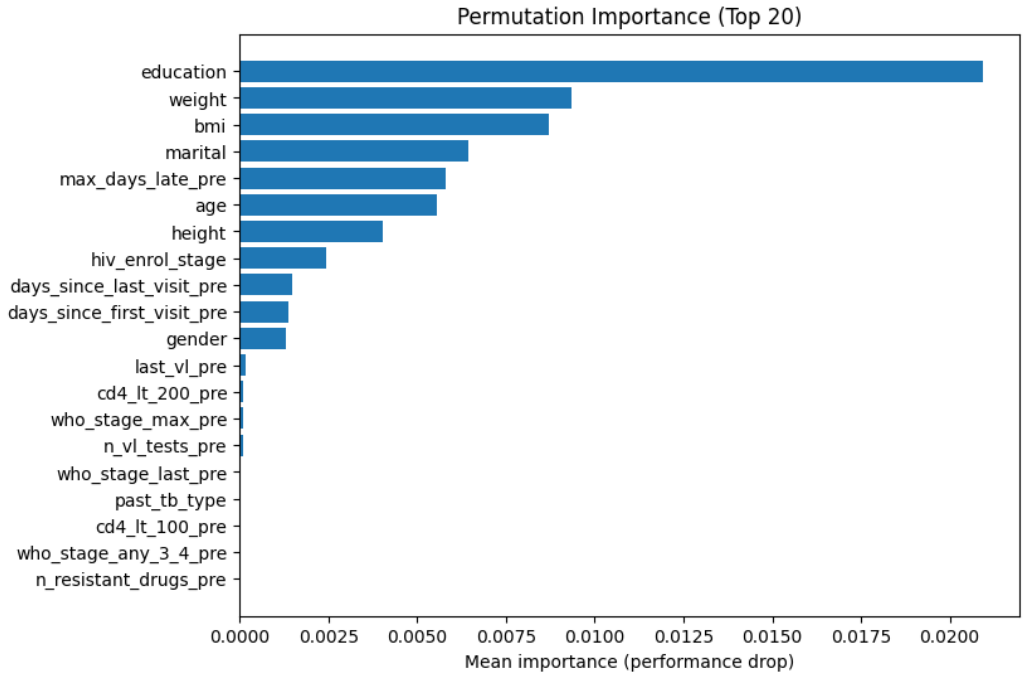

Figure L.6: *Permutation importance (top twenty features) for the IIT12m model on the held-out test set. Engagement features — mean and maximum days late, days since last and first visit — dominate, consistent with the operational definition of treatment interruption as a manifestation of disengagement from care. Anthropometric and demographic features contribute secondary signal.*

The IIT12m profile is the most operationally coherent of the three. Treatment interruption is, by construction, an engagement phenomenon, and the model's reliance on appointment-lateness and visit-recency features reflects exactly the signal a clinician would use to identify a patient at risk of disengaging. This coherence underpins the relatively high discrimination achieved for IIT12m (AUC-ROC 0.839) and supports the high-threshold, highly targeted operational protocol proposed for this outcome in Section 5.3.5.

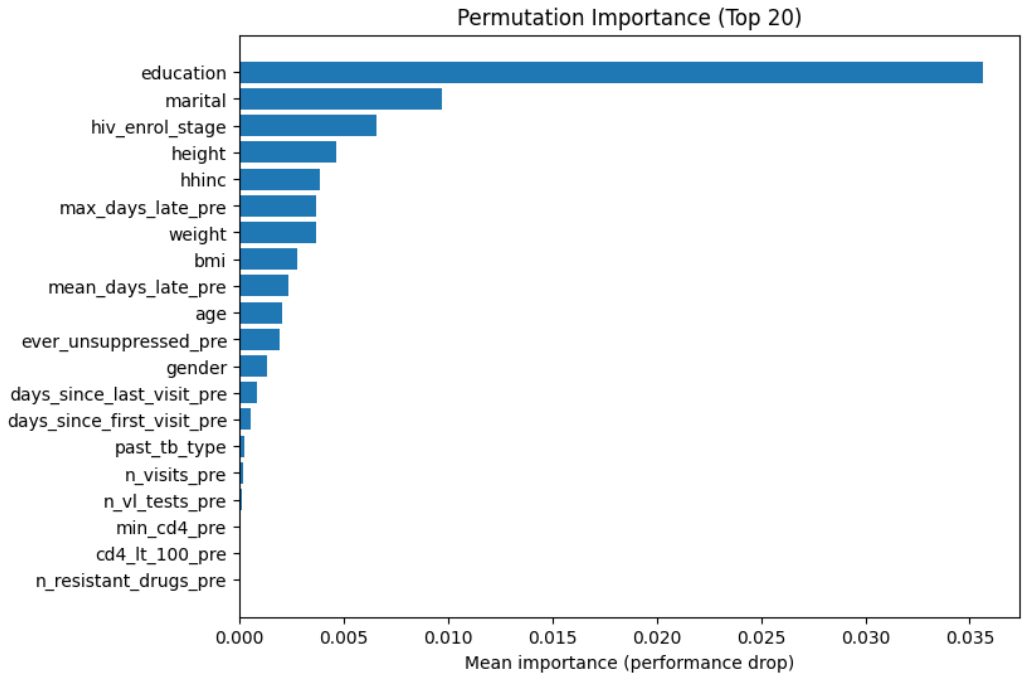

Figure L.7: *Permutation importance (top twenty features) for the UVL12m model on the held-out test set. Education level and its missingness indicator carry substantial weight alongside marital status, weight and age. The prominence of education — a field with only 4.7% completeness — is the empirical entry point for the missingness-outcome confounding analysis of Section 5.5.7.*

The UVL12m profile is the one that most demands cautious interpretation. Education level is both highly influential and severely incomplete, and the SHAP global summary for this outcome (Figure L.10) shows that the encoded absence of an education record, cat__education_None, is itself among the most influential signals. Whether this reflects a genuine health-literacy gradient in virological outcomes, an administrative pattern in which disengaged or harder-to-reach patients are both less likely to have education recorded and more likely to be unsuppressed, or a combination of the two, cannot be resolved from permutation importance alone. The missingness-outcome confounding test of Section 5.5.7 was designed precisely to interrogate this question, and the multilevel-modelling extension recommended by the CIDRZ AI Think-Tank (Appendix F) is the proposed route to disentangling the genuine signal from the structural artefact.

L.3 SHAP Global Summary (Beeswarm) Plots by Outcome

The SHAP global summary, or beeswarm, plot displays the distribution of SHAP values for each feature across all test-set patients. Each point is one patient; horizontal position encodes the signed

SHAP contribution (right increases predicted risk, left decreases it) and colour encodes the underlying feature value (red high, blue low). Features are ordered top-to-bottom by mean absolute SHAP value. These plots, reproduced as Figures 40 to 42, complement the permutation-importance rankings of Section N.2 by revealing not only how influential each feature is but the direction and value-dependence of its influence.

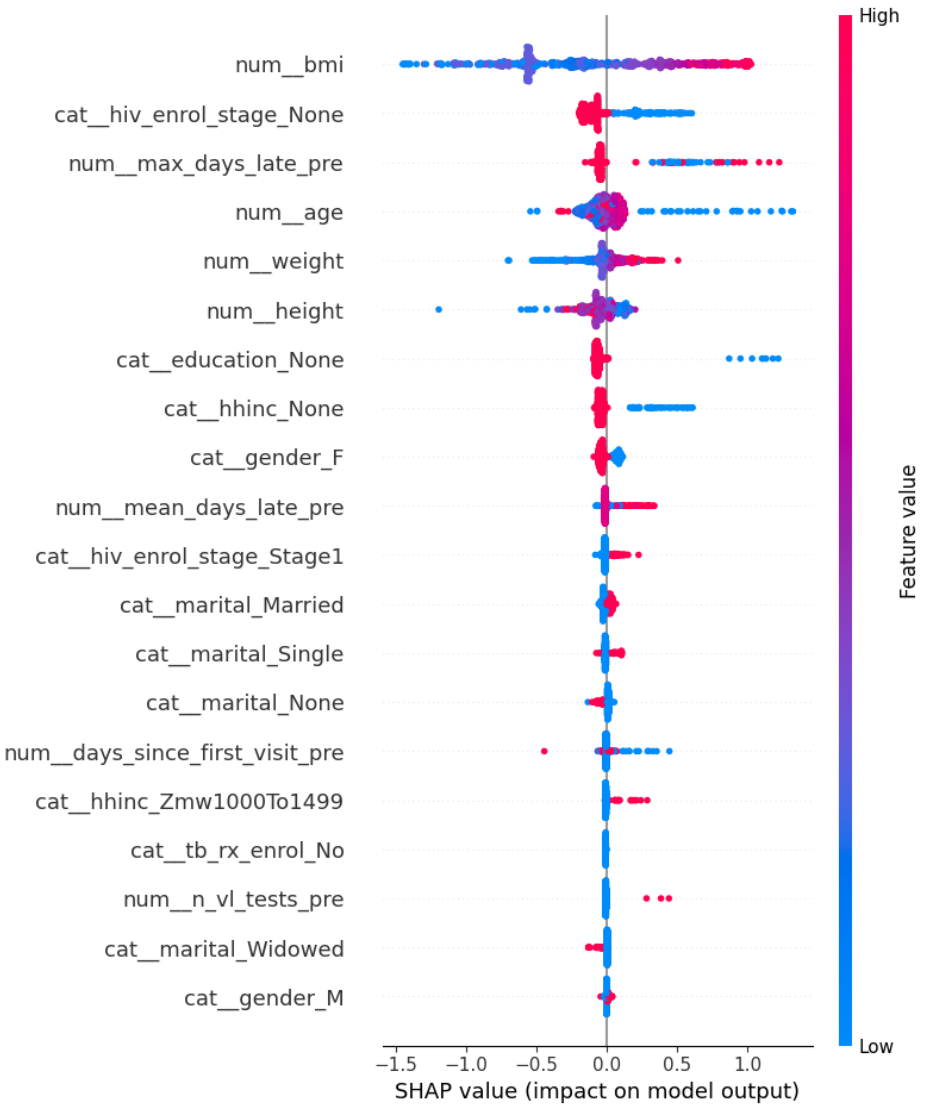

Figure L.8: *SHAP global summary (beeswarm) for the TB12m model on the held-out test set. BMI shows the widest and most clearly value-dependent spread: low-BMI patients (blue) cluster on the risk-increasing side and high-BMI patients (red) on the protective side, the expected direction. Missing HIV enrolment stage and extreme appointment lateness show distinct risk-increasing tails.*

The TB12m beeswarm confirms that the model's use of BMI is directionally correct and monotonic, which is an important sanity check: a model that assigned higher TB risk to higher-BMI patients would be clinically implausible and would signal a feature-encoding error. The clean separation of the BMI cloud by colour is therefore evidence in favour of the model's validity for this outcome. The risk-increasing tails of the missingness indicators (cat__hiv_enrol_stage_None, cat__education_None) are present but narrower than for UVL12m, consistent with the more clinical permutation profile of Figure 38.

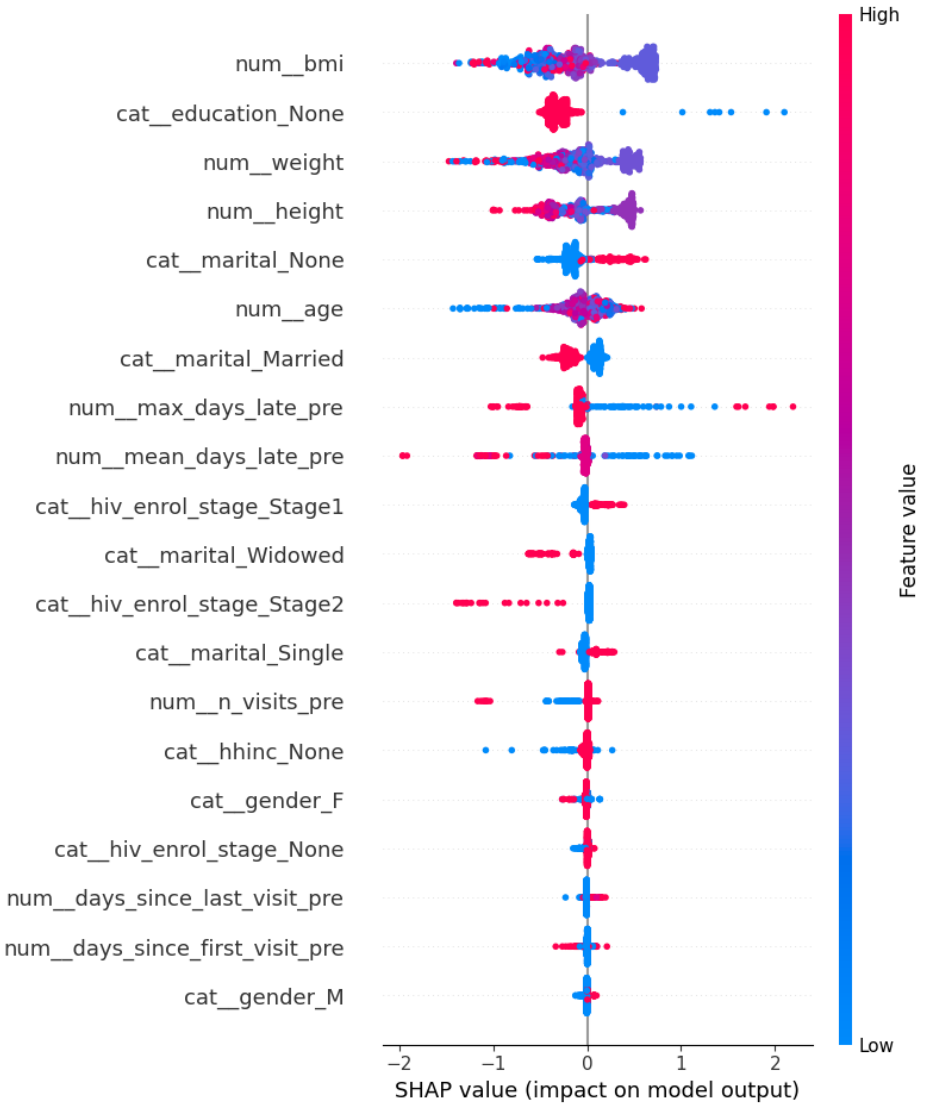

Figure L.9: *SHAP global summary (beeswarm) for the IIT12m model on the held-out test set. Engagement features dominate the upper rows. High values of maximum and mean days late (red) push predictions toward higher*

interruption risk, the operationally expected direction. The compactness of most feature clouds around zero reflects the extreme class imbalance of this outcome (0.63% positive in the test set).

The IIT12m beeswarm visually encodes the rare-outcome challenge: for the large majority of patients almost every feature contributes near-zero, and only the minority with adverse engagement histories receive substantial positive contributions. This is consistent with a model that is confidently identifying a small, well-characterised high-risk group rather than spreading modest risk across the whole population, which is the desirable behaviour for a targeted-intervention outcome and is reflected in the near-0.9 operating threshold adopted for IIT12m in Section 5.3.5.

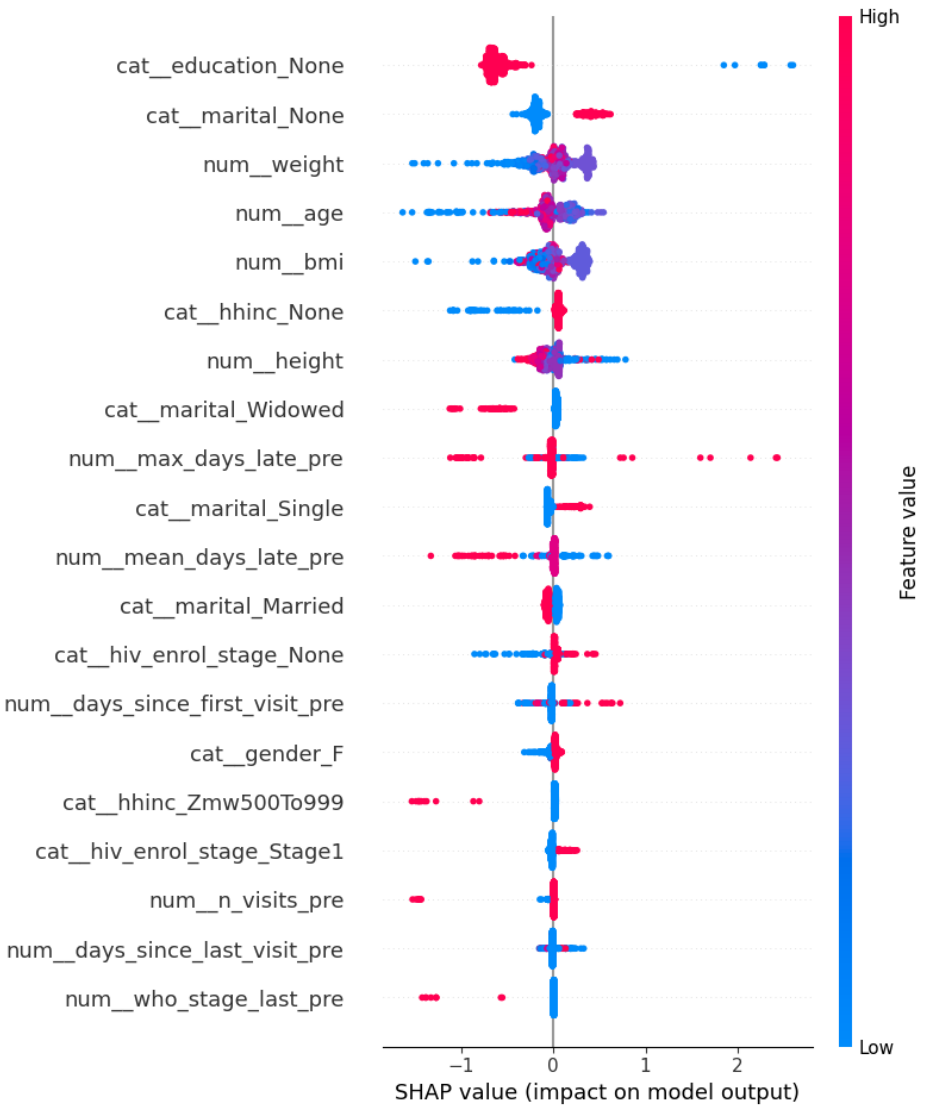

Figure L.10: *SHAP global summary (beeswarm) for the UVL12m model on the held-out test set. The encoded absence of a recorded education level (cat__education_None) is the top feature, and its value-dependence is the inverse of a clinical gradient: the small set of patients for whom education is recorded (blue) receive large positive or negative contributions, while the large majority with no record cluster tightly. This is the visual signature of the missingness-driven prediction pattern interrogated in Section 5.5.7.*

Figure L.10 is the single most important diagnostic figure in the thesis for the equity argument. The dominance of an encoded missingness indicator at the top of the UVL12m beeswarm, combined with the extreme single-feature SHAP contribution observed for the highest-risk patient (Figure L.1, cat__education_None = +2.58), demonstrates that for this outcome the model is, to a substantial degree, predicting on the basis of who has an education record rather than what that record contains. Because education completeness is lowest at the facilities serving the most deprived populations (1.2% at Matero Main against 7.9% at Kanyama, per Appendix D), a naive deployment that acted on these scores without safeguards would risk systematically mis-prioritising patients along lines correlated with deprivation. This is the empirical reason the proposed deployment architecture (Chapter 6) makes human-in-the-loop review mandatory for any prediction whose leading SHAP contributor is a missingness indicator, and the reason the governance architecture treats algorithmic-fairness audit as a model-approval gate rather than a post-deployment monitoring activity.

Taken together, the ten figures of this appendix constitute the reproducible evidentiary core of the empirical chapters. They show a data foundation that is systemically incomplete in precisely its most predictive fields (Section N.1); a set of models whose population-level reliance is clinically coherent for TB12m and IIT12m but partly missingness-driven for UVL12m (Section N.2); and a directional, value-dependent confirmation of these patterns at the level of individual SHAP contributions (Section N.3). The consistency of the story across the three analytical lenses — completeness diagnostics, permutation importance and SHAP — is what gives the thesis its central claim its empirical weight: that trustworthy artificial intelligence in resource-limited public health depends as much on the quality and governance of the underlying data as on the sophistication of the algorithm applied to it.

Appendix M: Glossary of Terms

This glossary defines technical and contextual terms used throughout the thesis, organised alphabetically.

Term	Definition
4IR	Fourth Industrial Revolution: the integration of digital, physical and biological systems characterising 21st-century industrial change.
ART	Antiretroviral Therapy: the combination of antiretroviral drugs used to treat HIV infection.
AUC-PR	Area Under the Precision-Recall Curve: a discrimination metric particularly informative for imbalanced classification tasks.
AUC-ROC	Area Under the Receiver Operating Characteristic Curve: a threshold-independent discrimination metric.
Brier Score	The mean squared difference between predicted probability and observed outcome; a standard calibration metric.
Calibration	The agreement between predicted probabilities and observed outcomes.
CDSS	Clinical Decision Support System: a software tool that supports clinical decision-making by clinicians.
CIDRZ	Centre for Infectious Disease Research in Zambia: a leading national health research institution.
DATIM	Data for Accountability, Transparency and Impact Monitoring: the PEPFAR reporting platform.
DHIS2	District Health Information Software 2: a national aggregate health information system.
DISA	A laboratory information management system used in Zambia for viral load and CD4 results.
DQA	Data Quality Assessment: systematic evaluation of data fitness for purpose.
DQI	Data Quality Index: a composite metric summarising data completeness across fields.
DSD	Differentiated Service Delivery: the operational paradigm of matching service intensity to patient risk.

Term	Definition
EHR	Electronic Health Record: a digital record of patient clinical information.
ELMIS	Electronic Logistics Management Information System: the pharmaceutical supply chain platform.
FHIR	Fast Healthcare Interoperability Resources: the HL7 standard for health data exchange.
HAART	Highly Active Antiretroviral Therapy: an earlier term for combination ART.
HL7	Health Level Seven International: the standards development organisation for health data exchange.
IIT12m	Interruption in Treatment within 12 months: a binary outcome operationalised in this thesis.
IPT	Isoniazid Preventive Therapy: TB prevention for PLHIV.
LMIC	Low- and Middle-Income Country: the World Bank classification used in this thesis.
MICE	Multiple Imputation by Chained Equations: a model-based missing-data imputation method.
ML	Machine Learning: the family of computational methods that learn patterns from data.
MoH	Ministry of Health: the Zambian government ministry responsible for health policy.
M&E	Monitoring and Evaluation: the systematic measurement of programme performance.
OpenHIE	Open Health Information Exchange: an international reference architecture for health data exchange.
OpenMRS	Open Medical Record System: an open-source EHR platform widely deployed in LMICs.
PEPFAR	United States President's Emergency Plan for AIDS Relief.
PHO	Provincial Health Office: the provincial-level administrative unit of Zambia's health system.
PLHIV	People Living with HIV.
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

Term	Definition
PROBAST	Prediction Model Risk of Bias Assessment Tool.
ROC	Receiver Operating Characteristic curve.
SHAP	Shapley Additive Explanations: a method for explaining individual ML predictions.
SmartCare	Zambia's national HIV electronic health record platform.
SDG	Sustainable Development Goal: the 2030 Agenda for Sustainable Development.
TB	Tuberculosis.
TB12m	Tuberculosis within 12 months: a binary outcome operationalised in this thesis.
TreeSHAP	A SHAP variant optimised for tree-based models with exact computation.
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis.
UNAIDS	Joint United Nations Programme on HIV/AIDS.
UVL12m	Unsuppressed Viral Load within 12 months: a binary outcome operationalised in this thesis.
WHO	World Health Organization.
XAI	Explainable Artificial Intelligence.
XGBoost	Extreme Gradient Boosting: a high-performance gradient boosting library.
ZCAS	Zambia Centre for Accountancy Studies University.
ZICTA	Zambia Information and Communications Technology Authority.