Out of the

Division of Infectious Diseases and Tropical Medicine, Medical Center of the University of Munich (LMU), Munich, Germany

HIV Viral load measurement in the Public Health Approach to HIV/AIDS - Developing a clinical score for patient management to identify patients at risk of failing HIV treatment.

Doctoral Thesis

for the awarding of a Doctor of Philosophy (Ph.D.)

at the Medical Faculty of

Ludwig-Maximilians-Universität, Munich

submitted by

Tessa - Suntje Lennemann

born in

Tuebingen

submitted on

31.10.2019

Supervisors LMU:

Habilitated Supervisor Prof. Dr. Michael Hoelscher

Direct Supervisor Dr. Arne Kroidl

Supervisor External:

Local Supervisor Dr. Omari Salehe

Reviewing Experts:

1st Reviewer Prof. Dr. Michael Hoelscher

2nd Reviewer Dr. Arne Kroidl

Dean: Prof. Dr. med. dent. Reinhard Hickel

Date of Oral Defense: 16. November 2020

Affidavit

Tessa Suntje Lennemann

Name

Street

Zip code, town

Country

I hereby declare, that the submitted thesis entitled

HIV Viral load measurement in the Public Health Approach to HIV/AIDS -

Developing a clinical score for patient management to identify patients at risk of failing HIV treatment.

is the result of my own work. I have only used the sources indicated and have not made unauthorised use of services of a third party. Where the work of others has been quoted or reproduced, the source is always given.

The submitted thesis or parts thereof have not been presented as part of an examination degree to any other university.

I further declare that the electronic version of the submitted thesis is congruent with the printed version both in content and format.

Bonn, 21.02.2021 — Tessa Lennemann

Place, Date — Signature of PhD Candidate

This work is dedicated to William Forrester Green, my dearly missed husband

Thank you for your faith in me and your never-ending love.

PhD Program Medical Research - International Health

CIHLMU Center for International Health

Ludwig-Maximilians-Universität, Munich

Title Page

HIV Viral Load Measurement In The Public Health Approach To HIV/AIDS -
Developing A Clinical Score For Patient Management To Identify Patients At Risk Of Failing HIV Treatment With An HIV Viraemia Above 400 Copies/ml In West Tanzania.

Key Words

HIV, ART, Program outcome, Virologic Suppression, Predictive Clinical Score

1 Abstract

Background

To end AIDS by 2030, the WHO 90-90-90 targets call for 90% virologic suppression in those on ART. In this context, it is crucial to understand which factors drive virologic suppression and how available resources can be targeted most effectively. This thesis evaluates a large HIV treatment program in Tanzania and explores performance and factors associated with treatment outcome on individual and on health system level. It then develops a clinical score to predict virologic failure and optimize patient management.

Design

This cross-sectional facility-based study assessed 702 patients stratified by time on ART at 7 study sites selected to represent regions of the study area and health care level.

Methods

Facility and patient-level information were collected during a single study visit. Logistic regression analysis and Generalized Boosted Model Technique derived Propensity Score Methods were used to explore health system and individual-level factors associated with virological failure. Predictive multilevel mixed logistic regression models were developed, externally validated and simplified into a normogram for the clinical score which was then tested against WHO recommended failure criteria using Decision Curve Analysis.

Results

Within the population on ART, 89% was virologically suppressed below 1000 copies/ml and 86% below 400 copies/ml. Differences could be found between health care levels but not regions. The study site had a direct impact on treatment outcome on the individual and health system level. Performance of the clinical scores was high with a ROC-AUC of 0.8 in the training, and ROC-AUC between 0.7 and 0.8 in the population and the geographic validation dataset. Decision Curve Analysis showed a net benefit against the WHO routine and targeted viral load monitoring strategies.

Conclusion

To fully reach the “the last 90” health system-level interventions should support sites. On individual level, the clinical score developed could be used to better identify and manage individuals at risk of treatment failure.

2 Table Of Contents

3 List Of Figures

4 List Of Tables

5 Abbreviations

AIC	Akaike Information Criterion
AIDS	Acquired Immunodeficiency Syndrome
ALT	Alanine Aminotransferase
ART	Antiretroviral Therapy
ATE	Average Treatment Effect
BMI	Body Mass Index
CI	Confidence Interval
C-statistics	Concordance Statistics
CTC	Care and Treatment Centre
dL	Decilitre
EFV	Efavirence
FTC	Emtricitabine
GBM	Generalized Boosted Model
Hb	Haemoglobin
HCW	Health Care Worker
HH	Household
HIV	Human Immunodeficiency Virus
IRS	Immune Reconstitution Syndrome
IS	Immunological Success
Kg	Kilogram
KS	Kolmogorov Smirnov Statistic
LFT	Liver Function Test
MCV	Mean Corpuscular Volume
MbDH	Mbeya District Hospital
MbRH	Mbeya Regional Hospital
MHRP	Military HIV Research Program
MZRH	Mbeya Zonal Referral Hospital
NVP	Nevirapine
NNRTI	Non Nucleoside Reverse Transcriptase Inhibitor
NRTI	Nucleoside Reverse Transcriptase Inhibitor
OPD	Outpatient Department
p	Significance test result using z statistics
PCP	Pneumocystis jiroveci Pneumonia

PEPFAR	US President's Fund for Emergency AIDS Relief
PPS	Proportional Probability Sampling
PS	Propensity Score
RBC	Red Blood Count
ROC-AUC	Area Under The Receiver Operating Characteristic Curve
RuDH	Rukwa District Hospital
RuRH	Rukwa Regional Hospital
RvDH	Ruvuma District Hospital
RvRH	Ruvuma Regional Hospital
SDG	Sustainable Development Goal
SMD	Absolute Standardized Mean Difference
TB	Tuberculosis
TDF	Tenofovir
TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
UN	United Nations
UNAIDS	United Nations Programme on HIV/AIDS
VS50	Virologic cut-off above/below 50 copies/ml
VS400	Virologic cut-off above/below 400 copies/ml
VS1000	Virologic cut-off above/below 1000 copies/ml
WBC	White Blood Count
WHO	World Health Organization
WRSHCP	Walter Reed Southern Highlands HIV Care Program
µL	Microliter
VF	Virologic failure
VS	Virologic suppression
VS1000	Virologic suppression below 1000 copies/ml
VS400	Virologic suppression below 400 copies/ml
VS50	Virologic suppression below 50 copies/ml

6 Introduction

Since more than a decade, coping with the Human Immunodeficiency Virus (HIV) epidemic has been in the centre of global public health activities. The introduction and increasing availability of antiretroviral therapy (ART) has substantially transformed HIV related interventions: while initially HIV therapy was delivered purely for humanitarian reasons, it has become central to reach the United Nations Sustainable Development Goal 3 (SDG3) which envisions to “end AIDS by 2030”[1] as one of the health-related objectives. The 90-90-90 strategy of the World Health Organisation (WHO) guides the concerted efforts of national and international partners towards SDG3: the first two “90” stand for reaching 90% of individuals living with HIV to let them know their status and linking 90% of those reached to treatment and care. The “last 90” commit to achieving serologic virologic suppression in 90% of those receiving ART. It is envisioned that with this strategy, the rate of new infection can be reduced to below 1:1000 per annum, which is considered the threshold signifying epidemic control.

Virologic suppression can be described as the outcome of a complex and dynamic interplay between biological factors driving the host-virus interaction and factors on individual and community level that impact serologic ART concentrations which halt viral replication. In such a concept, the health system would be an important community-level factor impacting treatment success. The “last 90” thus are not a static end result of a public health intervention, but the expression of a fragile balance between favourable and unfavourable factors within the unit of observation which can be it an individual, a region, a time period or a population.

To achieve “the last 90” as a contribution towards ending AIDS in 2030, virologic suppression has to be not only achieved but sustained within programs, across populations served, on sub-national level and overtime while maximising the impact of available resources through their targeted use.

In Sub-Saharan Africa, 90-90-90 is implemented through national governments supported by international implementing partners who augment and complement national structures and services relevant for HIV management. The Walter Reed Southern Highlands HIV Care Program (WRSHCP) which is evaluated in this thesis exemplifies such a collaboration between national and international actors, supporting ARV delivery of the governmental health system in western Tanzania.

For such a program, it is not only important to understand how far it has been able to achieve 90-90-90 on the programmatic level, but also how it can best improve equitable service delivery. In this respect, interventions to improve quality of care could strengthen regional service, health care levels or specific sites, but interventions could also be targeted to the characteristics of the individual for example through a predictive clinical score that could help identify individuals with a higher risk of failing treatment. Individuals identified through such a score could then be linked to additional targeted services. Especially with the increasing introduction of digital patient management systems in public health programs, a predictive score could be integrated into patient management software and the predicted risks could initiate a triage system that could fast-track low-risk individuals while focussing resources on those most likely to benefit.

To contribute to this understanding, the current thesis addresses the following research questions, using the data generated through a multi-centre cross-sectional cohort study in the population reached by the WRSHCP:

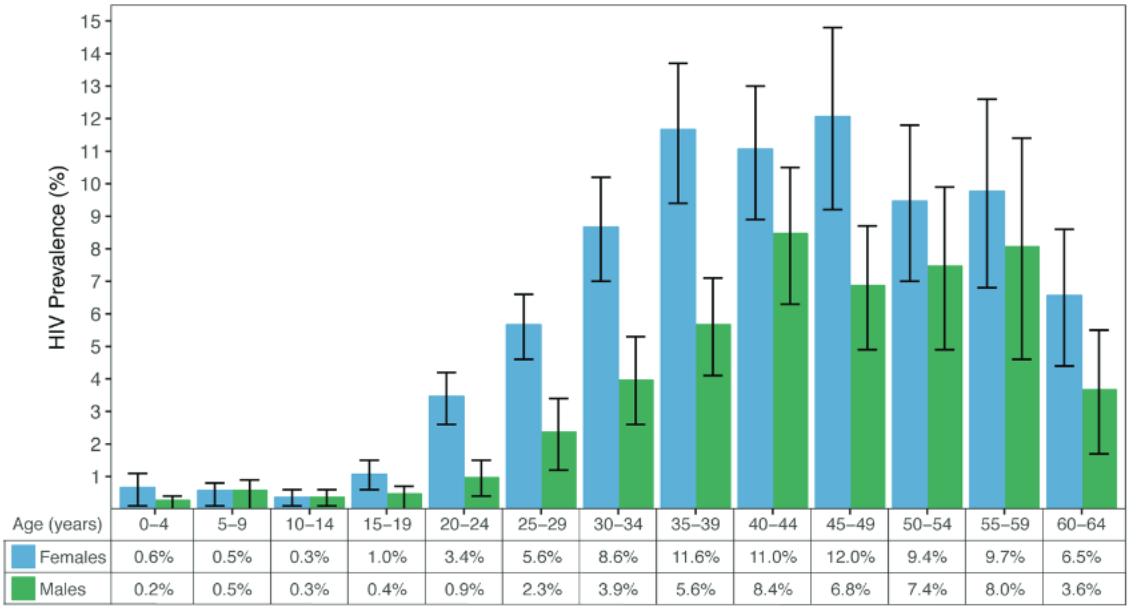
1) Does the national ART program supported by WRSHCP achieve “the last 90” on the programmatic level?
2) Can differences in virologic suppression be observed that are associated with regions, health care levels or individual Care and Treatment Centres (CTCs)?
3) Does programmatic outcome differ in subpopulations defined by time on treatment and which other factors are associated with virologic failure on the level of an individual accessing ART within the program?
4) Can a predictive diagnostic score be developed that could identify individuals with a higher risk of exhibiting virologic failure?

As the WRSHCP is implemented through the national structures, results generated on program level can be generalized to the national ARV service delivery in the region covered by the program and further inform other ARV programs in comparable settings and with comparable populations. This transfer will be supported by the use of three different definitions of virologic failure as an outcome measure based on the rationales presented in 6.1.3.

6.1 Background

6.1.1 Epidemiology Of HIV/AIDS In Tanzania

After the first three cases of AIDS were reported from Northern Tanzania in 1983, HIV infection had spread across the country by 1986. By 2017 the national prevalence was estimated to be 5% [2] and has since then plateaued at this level, with 5% national prevalence reported in the most recent national survey conducted in 2018 [3]. As illustrated in Figure 1, the population is not uniformly affected. Women have an overall higher prevalence than men (national prevalence of 6.5 per cent among females and 3.5 per cent among males) [3] and contract HIV earlier in life. While female peak prevalence of 12% is found at 45 to 49 years, the prevalence in males is half that in women during adolescence and as young adults and only peaks between 40 and 44 years at 8.4% [3]. Annual HIV incidence in the overall population is estimated at 0.3%, with 0.4 and 0.14 in women and men respectively [3].



Age (years)	0–4	5–9	10–14	15–19	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59	60–64
Females	0.6%	0.5%	0.3%	1.0%	3.4%	5.6%	8.6%	11.6%	11.0%	12.0%	9.4%	9.7%	6.5%
Males	0.2%	0.5%	0.3%	0.4%	0.9%	2.3%	3.9%	5.6%	8.4%	6.8%	7.4%	8.0%	3.6%

(error bars represent 95% confidence intervals)

Figure 1: HIV Prevalence by Gender and Age 2017 [3]

Sub-national differences exist in HIV prevalence, with lower prevalence along the coastal regions and a concentration of the epidemic in the highlands of Tanzania where this study was conducted. The study regions Ruvuma, Mbeya, Katavi and Rukwa are above the national average with an HIV prevalence of 5.6, 9.3% 5.9% and 4.4% respectively (Figure 2) [3].

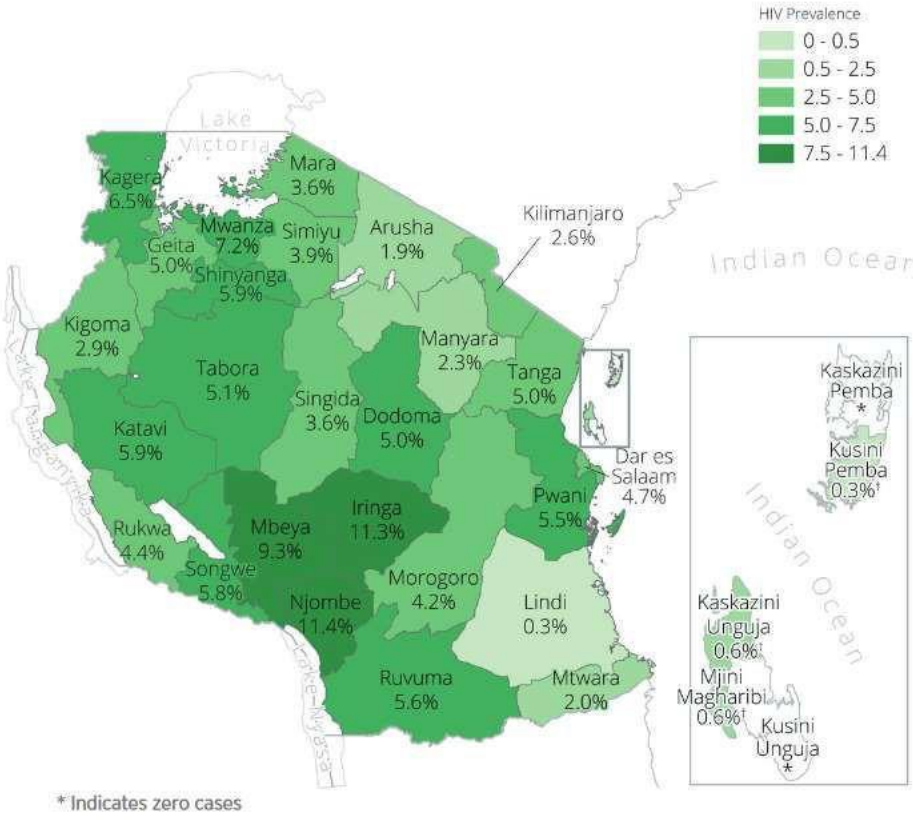


Figure 2: HIV Prevalence by Region 2016 [3]

6.1.2 National And International Response To HIV/AIDS In Tanzania

The initial response to HIV/AIDS in Tanzania focused primarily on prevention, but in 2001, the national HIV/AIDS policy acknowledged the right to HIV treatment for people living with HIV [4]. First treatment targets were set in the national strategic plan for HIV/AIDS 2003-2008, which aimed to accelerate ART treatment access from 16.000 people on ART in 2004 to 423.000 in 2008 [5]. In the following National Strategic Framework 2008-2012, access to treatment and care had become one of three priorities [6]. The third strategic plan for the period of 2013-2017 fully centred around comprehensive antiretroviral therapy, considering the continuum of care from HIV testing to ART access the most important area of primary investment [7].

In 2014, the Joint United Nations Programme on HIV/AIDS (UNAIDS) launched the ambitions 90–90–90 targets that aimed to give 90% of all People living with HIV knowledge of their HIV status, 90% of those diagnosed HIV sustained antiretroviral therapy and 90% of all people receiving ART to be virologically suppressed. Tanzania incorporated 90-90-90 as a pathway to reach the “three zeros” – Zero New

HIV Infections, Zero AIDS-related Deaths, and Zero Stigma and Discrimination –at the heart of the 2013-2017 strategic plan [7], reflecting the countries' commitment to HIV elimination. It is estimated that for every 1% increase in ART coverage, an estimated 1-2% decline in HIV transmission risk can be expected and that a successful 90-90-90 strategy will allow achieving virological suppression in 73% of the total HIV infected population [8]. Such a level of suppression is expected to result in epidemiological control of HIV at the population level, which is commonly defined as HIV transmission below 1:1000 per annum.

The increased focus on treatment provision resulted in an acceleration of treatment access and in 2016 an estimated 62% people living with HIV (850.000 [CI 480.000 – 740.000]) of those in need of treatment were on ART [9]. At the same time, the increase of the CD4 threshold for ART start from 250 to 500 CD4 cells and later to the "test and treat" policy led to an increase of the overall population of HIV infected individuals eligible for treatment [8,10,11]. However, these changes also increased the pressure on infrastructure and patient load at the facilities at a time when funds for HIV specific programming stagnated as donors re-oriented themselves to the new UN Sustainable Development Goals [12] that prioritize health system support over support for disease-specific interventions [13,14].

The 2016/2017 Tanzania HIV Impact Survey (THIS) provides information on the progress made 3 years after the beginning of 90-90-90: It found 52.2% of PLHIV aged 15 to 64 years knew their HIV positive status, and among these, 90.9% were on ART. Of those on ART, 87.7% (89.2 per cent of HIV-positive females and 84.0 per cent of HIV-positive males) were virally suppressed below 1000 copies/ml [3]. However, due to the large proportion of undiagnosed HIV infections, the goal of 75% viral load suppression on a population level could not be achieved [3]. Differences in suppression could be seen by age and gender, and gender disparity in viral load suppression was greater at a younger age [3]. Additionally, regional variations in the viral load suppression were observed, with 57.4% and 56.7% for Mbeya and Ruvuma Region, while Rukwa and Katavi region (which both were unified in one district when this study was conducted) showed an even lower suppression of 42.9% and 47.3% respectively (Figure 3) [3].

Despite the nearly universal adoption of the 90-90-90 strategy in Tanzania as much as internationally, it's practical implementation and the possibility to achieve the

underlying objectives have been repeatedly questioned, arguing that even in high-resource countries, 90-90-90 has not been achieved [15].

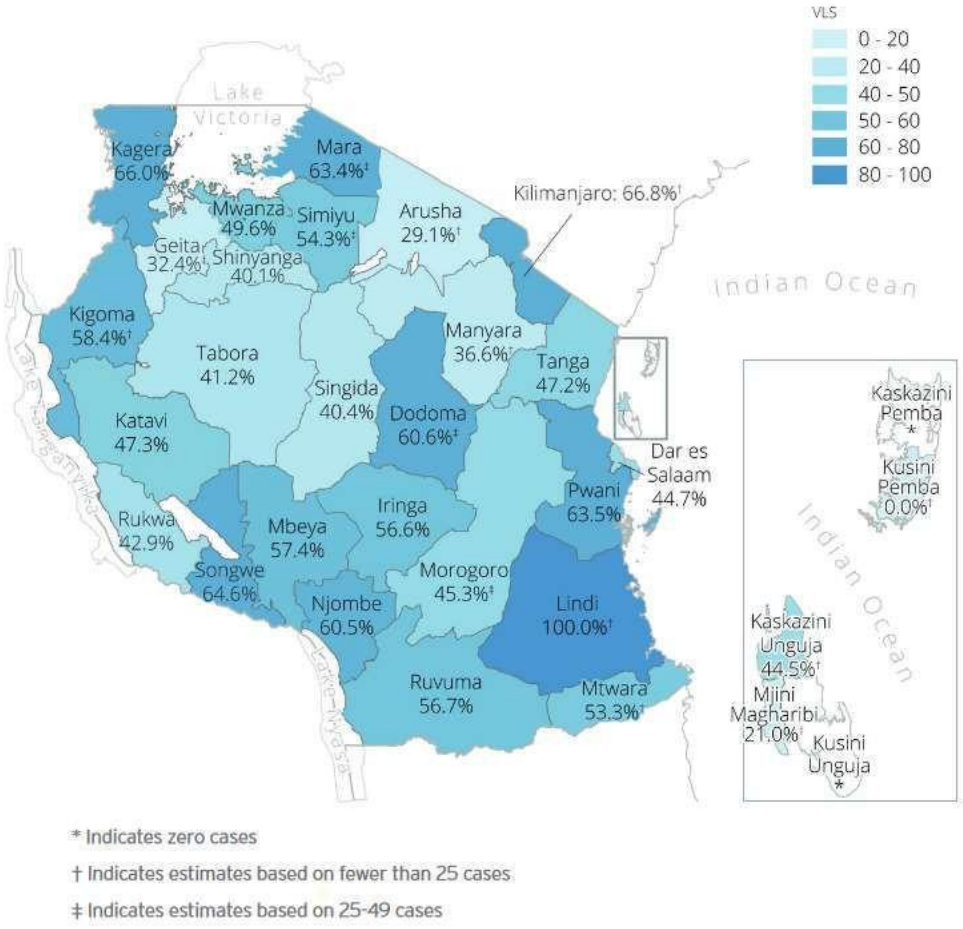


Figure 3: Tanzanian HIV Impact Survey 2016-2017 (THIS): Virologic Suppression <1000 copies in HIV Infected Individuals by Region [3]

The Walter Reed Southern Highland Care And Treatment Program In Tanzania As Example For Implementation of 90-90-90 In Tanzania

The Walter Reed Southern Highland Care and Treatment Program in Tanzania (WRSHCP) is an example of concerted national and international efforts to control HIV in Tanzania and is the program evaluated in this thesis:

The President's Emergency Plan for AIDS Relief (PEPFAR) invested a total of 4,205,708,000 USD between 2004 and 2017 into HIV related interventions [16], with the Walter Reed Army Institute of Research being one of its prime recipients. This Institute supports a network of sites conducting HIV prevention, care, and treatment as much as HIV vaccine and therapeutics research across more than four other

Africa countries [17] and started to support the governmental roll-out of ART in Tanzania in 2004 in the context of the larger United States Embassy PEPFAR mission in Tanzania. In this context, the Walter Reed Southern Highland Care and Treatment Program (WRSHCP) was conceived with a mission to develop and implement a comprehensive community approach to HIV care and treatment in Mbeya, Ruvuma and Rukwa Region. Activities by WRSHCP are conducted in collaboration with the Mbeya Zonal Referral Hospital (MZRH) and the Mbeya, Rukwa and Ruvuma Regional Medical Offices in close coordination with the Ministry of Health and the National AIDS Control Programme. WRSHCP provides technical assistance for clinician and laboratory training using the national curriculum as well as support for laboratory services. It thus supports the facility-based national health system on referral, regional and district level.

6.1.3 Measures of Outcome To Determine Treatment Success

The introduction of 90-90-90 also shifted the outcome measures used to judge treatment success on the project and individual level. On the individual level, clinical or immunological parameters were replaced by viral load measurement [18, 19], while on the population level, the focus was shifted from the number of people accessing ART to a proportion of the client population with viral suppression as a programmatic outcome measure.

Blood viral HIV concentration has become the standard to judge treatment effectiveness as much as transmission risk and is the unit within which treatment success is measured. But the cut-off to convert the continuous viral load measurement to a binary measure of success or failure depends on underlying aims and settings, with the following three main cut-offs commonly used internationally:

<u>Below 1000 Copies/ml – Definition Of The World Health Organisation (WHO)</u>

Current WHO guidelines use a cut-off of above 1000 copies/ml as the definition of virologic failure both for defining treatment failure in the individual [20-22] and on programmatic level [23]. Individuals who fall below this cut-off are counted under the “last 90”. The choice of this cut-off is based on the following considerations:

1) The cut-off is feasible in a setting with constrained infrastructure.

The higher cut-off was chosen for considerations of feasibility in low resource countries as it would allow the use of dried blood spots - a sampling method already

established for early infant HIV diagnostics - in the context of treatment monitoring, simplifying sample generation and transportation for the price of a higher detection limit for HIV in these samples [24].

2) The cut-off prevents unnecessary treatment switches without compromising clinical prognosis for People Living With HIV:

Temporary viremia under treatment is frequently observed: next to drug failure due to resistance development, co-infections, reduced adherence, co-medications and increased viral production in compartments other than blood [25] can all contribute to low-level replication or temporary “blips” in the presence of effective treatment that is not always associated with resistance development [26, 27]. Without the opportunity to conduct a drug resistance test and acknowledging the large intervals between the viral load tests arising from the prohibitive high costs of the viral load tests, a lower cut-off as a definition of treatment failure - such as >50 copies used in high resource countries - would result in a high level of unnecessary treatment switches to second-line therapy, compromising cost-effectiveness of the HIV program as a whole and unnecessarily limiting future treatment options for the individual. As viral loads below 1000 copies/ml have been associated with low risk of disease progression [28], this cut-off is thus considered a suitable threshold where clinical morbidity and mortality of HIV infected individuals is not compromised.

3) The cut-off is relevant with respect to HIV transmission

On population level under the aspect of treatment as prevention, HIV transmission has been shown to be dependent on the level of viremia [29] and is considered unlikely below 1000 copies/ml [30]. In serodiscordant heterosexual couples in Uganda, transmission probabilities significantly dropped from 0.0023 per act at 38500 copies/mL to 0.0001 per act at viral loads of less than 1700 copies/mL [31]. Also in respect to mother to child transmission of HIV, only 1% vertical transmission has been reported under this cut-off [32]. In consequence, this cut-off is seen as a pragmatic compromise that is feasible in a real-life setting.

<u>Below 50 Copies/ml – The Linear Detection Limit Of Commonly Used Viral Load Tests</u>

The WHO definition of treatment failure of viral load above 1000 copies/m has been contested in various ways: Concerns were raised about applicability across different

populations such as infants, children, adolescents and pregnant and breastfeeding women [30] and the safety of this rule for individuals with low-level viremia below the WHO threshold. Especially prolonged low-level viremia has been linked to subsequent failure in high-income settings [27, 33-35] and individuals with replicating virus have been shown to have a higher risk of morbidity and mortality independent of their immunological status. Similar evidence is emerging from Sub-Saharan Africa: Studies assessing resistance patterns in patients failing with low-level replication indicate the substantial presence of drug resistance that compromises future treatment options and argues for a lower cut-off for individual client management [35, 36]. A large South African cohort managed according to WHO guidelines most recently reported 23.3% low-level replication, with an increased risk of treatment failure or a switch to second-line treatments with increasing viral load. The study reported a hazard ratio of 3 for replication under 200, 4.2 for clients replicating between 200 and 400 copies/ml and 7.7 for those between 400 and 1000 copies/ul [37, 38]. These considerations support the use of this lower cut-off of 50 copies/ml or below and drug resistance test driven treatment switch which is the gold standard in resource-rich settings.

Below 400 Copies/ml - A Cut-off For Quality Assessment Of Service Providers

A threshold of 400 copies – which was the detection limit of the first HIV viral load test available in the 1980ies- has been used alongside 1000 copies to describe treatment outcome especially of programs where repeated measurements were not feasible to avoid misclassification of transient viremia as treatment failure. In research-rich settings, >400 copies or > 200 copies are applied as cut-offs in the context of quality assessment of the health care service rather than individuals, adjusted by a case-mix that accounts for morbidity and mortality [39]. The Ryan White HIV/AIDS Program, which provides HIV primary medical care and support for uninsured people living with HIV in the United States by funding states, cities/counties, and local community-based organizations has included this cut-off for monitoring funded site performance in the context of it’s Performance Measure Portfolio [40].

Program Attrition As Form Of Treatment Failure

While treatment success can best be defined through viral load below a certain cut-off as outlined above, treatment failure does not only include viral replication above

this threshold in individuals remaining in care, but also a complete drop-out of the individual from the service. Real attrition may be due to self-transfer, death, true service drop-out, but on the population level, challenges related to documentation such as duplication of hospital IDs or dual registration of clients [41-44] can inflate patient numbers and thus bias results. Throughout public health programs in Africa, attrition is high, with an overall estimate of 33% attrition of which 18.8% are attributed to true treatment termination and 14.7% to mortality after 5 years on ART[45]. WHO has established <20% attrition of a program as a quality measure, but reports on attrition show a wide range [46] and attrition based on programmatic data is increasingly considered to be prone to overestimation as individuals might choose to self-transfer or re-access treatment following a treatment interruption[45, 47, 48] . Reasons for attrition also change over time, with the declining risk of death and increased undocumented transfers and treatment interruptions with longer time on treatment [43, 49] and may differ in urban and rural areas [43]. While the dynamics of attrition are poorly understood, it is likely that dynamics driving viral replication are also underlying attrition and identification of individuals likely to fail treatment might also help reduce real project attrition.

6.1.4 Factors And Dynamics Determining Viral Replication And Virologic Suppression

Viral replication in this thesis is seen as the product of complex dynamics and interactions between biological, socio-cultural and community-level factors that either impact drug levels hampering viral replication or affect the virus-host interaction driving HIV associated morbidity and mortality.

Individual-level Characteristics Impacting Adherence

On the level of an individual living with HIV, individual characteristics can impact the ability to adhere to the daily drug regimen, resulting in irregular dosing intervals, inadequate drug intake and drop-out from the ART program while individual biological factors - especially co-morbidities and their treatment - can interfere both with the ability of the individual to adhere to treatment as with the efficacy of the drugs through drug-drug interaction or compromised resorption. Unchecked viral replication, on the other hand, drives chronic inflammatory processes of immune activation, that in turn increase the risks for non-infectious co-morbidities such as cardiovascular diseases.

Among demographic variables that have been associated with treatment outcome in Tanzania are younger age [50] and male gender [49, 51, 52], marital status, level of education, literacy and distance to an ART treatment clinic. Many of these are also indicators of the economic and social status of an individual which in turn may impact individual ability to manage consequences of infection from stigma to practicalities surrounding clinic access and regular drug intake: Out of pocket payments are known barriers to health care access and are especially relevant in a disease that requires time-consuming life-long clinic visits on a monthly basis. Consequently, time and money available to access ART have been an established variable defining treatment failure [53]. On the other hand, the ability to manage stigma associated with HIV is crucial for continuous adherence and might be decreased with increased income and social standing especially in the public health HIV programs where confidentiality sometimes is constrained. As in many other diseases where no cure is offered through biomedicine, alternative health belief models are consulted by many clients to varying degrees next to the biomedical healthcare system [54, 55]. The personal concept of health and illness and the reliance on biomedical or alternative healing traditions will shape the way individuals interact with the biomedical treatment strategy provided by the health system and thus impact the virologic outcome.

Client-level factors interfere with the efficacy of ART, or by influencing viral replication through impacting immunological control. A variety of laboratory and clinical parameters has been associated with treatment outcome in resource-limited settings, of which the WHO clinical and immunological failure criteria are the best-established examples [51]. However, due to their low sensitivity and specificity of these clinical parameters, other variables are constantly explored as alternative surrogate parameter for treatment monitoring or to predict treatment failure: next to variables at treatment start, such as CD4 count, weight and poor functional status [46], a positive association with virologic suppression has been reported for haemoglobin concentration [53], total lymphocyte count [53] and CD4 counts at the time of assessment [46].

Further, the drug regimen provided in the context of first-line ART unsurprisingly is highly relevant for treatment outcome and defines the type and speed within which drug resistance emerges. In a comparable study conducted in Uganda, virologic suppression was positively associated with Efavirenz use [53]. Drug regimen also impact adherence as side effects often cause drug avoidance [56] either through

directly affecting drug intake for example when causing vomiting or confusion, through making the infection visible as in case of lipodystrophy or because the fear of side effects outweighs the expected or perceived benefit of the medication.

Drug-drug interactions of ART with co-medication further may complicate drug efficacy and tolerability, while prophylactic short-term use of ART – such as single-dose NNevirapine during pregnancy to prevent vertical transmission - might have an impact on the outcome of later ART due to long half-life and low resistance barrier of some drugs.

Community And Context Within Which HIV Treatment Takes Place

On a community level, dynamics driven by the environment within which treatment takes place can support or hamper individual treatment outcome: Client-Community interactions such as gender and age normative behaviour or economical discrepancies can impact individual adherence while cultural believes and norms modulate disease perception that guides individual health-related behaviour.

A functional personal social environment with strong social support networks has been shown to benefit treatment outcome [57]. Social support may mediate pressures and practical challenges that are a consequence of HIV infection and ART treatment. A protective factor in this respect, for example, is the ability to disclose one's own HIV infection within the family and to sexual partners so that the day-to-day reality of taking medication and attending CTCs is not complicated by the need of confidentiality. On a different level, responsibility for other individuals, be it in the form of their financial dependency, care for someone with HIV [57], or parenthood often serves as a motivation to adhere to ART [57]. Likewise, being employed has shown to be an international predictor of treatment adherence [58]. Support groups and treatment buddies can be counted in this context as an extended private social network.

Health System Impacting Treatment Outcome

Integrated into the various dynamics of the community, the health system plays an important role in mediating treatment outcome beyond providing the correct drugs consistently: Service satisfaction and client-centred service delivery define success at the client-clinic interface, while treatment can be compromised by facility-related

obstacles such as drug stock-outs, doctor to client ratios and other variables of service quality:

Complex dynamics at the client-clinic interface may impact individual adherence to ART [59] at various levels: Logistical aspects of clinic access such as distance from home to CTC, duration of the clinic visit and related monetary, time and logistical costs have been established factors influencing the individual ability to sustain ARV [57]. Long waiting time, friendly staff, level of confidentiality or supportive opening hours further can modify the interaction of the individual with the health system and impact willingness to prioritize regular treatment access over competing priorities in the individual's life. ART clinics may employ various measures to mediate at the client-clinic interface: telephone calls or automated SMS messages and home visits of outreach programs have shown a positive impact on retention in care [43]. Services provided at the clinic such as frequency of counselling, facility-based support groups or electronic reminders for the daily drug dosing [60, 61] but also complementary services such as nutritional support [62] are associated with positive treatment outcome, as is the wider clinic infrastructure such as availability laboratory reagents, drugs or space and overall years of clinic operation [51]. In this respect, Health Care Workers (HCW) are central to treatment outcome which is positively associate with HCW training levels [63, 64], task shifting, the use of peer educators, and HCW attitudes and resulting practices [65].

Clinic effectiveness may vary and clinics and might have the potential to maximize output with the existing infrastructure through streamlining workflow, roles and responsibilities. Di Giorgio et al. estimated that in Kenya, Zambia and Uganda facility capacity to initiate new individuals on ART could be increased by 40% if effectiveness could be raised from the below 50% observed to 80% [66] without introducing differentiated care models that may shift ART treatment and care to the community level.

Next to improving efficiencies of clinics, differentiated care models have been developed that diversify the health system through the addition of service delivery models that go beyond ART provision in a dedicated CTCs: Facility-based outreach services [46], community-based ART[65, 67, 68] or HIV services integrated into other hospital-based services such Outpatient Departments (OPD), Tuberculosis (TB) services or ANC services [62] have shown to positively impact identification, linkage

or retention and are considered efficient and cost-effective approaches that diversify the entry point to HIV treatment and care and address obstacles to treatment access. Further, as a positive impact on population-level disease control is only possible if service equity can be achieved, these models might better target the needs of specific sub-groups of the client population such as clinically ill or second-line patients, pregnant women [62, 67], HIV infected adolescents [50] or key populations [50] that otherwise might be left behind, jeopardizing the overall program outcome. Differentiated Care Models further reduce overall client volume at clinics as they shift ART services to the community level. This decongestion at the clinic level in itself has been directly associated with improved treatment outcome [69]. Nevertheless, differentiated care models mainly address the first and second 90, and especially in Tanzania CTCs remain the major providers of ART and thus central to reaching the “last 90”.

Quality of health care service has been directly linked to treatment outcome also in resource-rich settings: In 2010, national HIV/AIDS performance measures for system-level quality improvement were published for the US [70], and >80% of those measures received was associated with survival prior adjustment for disease severity [85]. Podlekareva used a similar set of indicators to assess the quality of care in the EuroSIDA study by comparing services provided by the study sites to the respectively applicable treatment guidelines. Regional differences observed could be correlated with the client-level virologic outcome. Nevertheless, it is unclear which parameters best describe the quality of care of a facility, and different outcome measures, such as viral suppression or retention in care, have shown to be only poorly correlated with each other.

6.1.5 Management Of Virologic Failure In The Public Health Approach

Virologic suppression under ART is desirable for the HIV infected individual as it prevents morbidity associated with HIV infection on the individual- and HIV transmission on the population level. Viral replication, on the other hand, can indicate incomplete drug adherence and increased risk of drug-resistant development.

WHO recommends two strategies depending on available resources to address virologic failure in individuals: When choosing routine viral load monitoring, viral load is monitored in all patients on ART at 6 and 12 months of after ART start and annually thereafter. Alternatively, a targeted viral load approach only in clients with

suspected treatment failure – such as those who meet the definition of clinical or immunological failure - is recommended. In any case, if the first viral load test shows a viremia above 1000 copies/ml, intensified adherence support interventions should be initiated and a second viral load should be performed after 2-3 months. If this second viral load remains above 1000 copies, treatment should be switched to second-line therapy [71]. However, due to delays surrounding sample transportation, availability of reagents and result feedback, this process often takes substantially longer [72] and the complexity of this algorithm is considered one of the reasons why uptake of second-line ART is much lower than what would be expected, resulting in preventable morbidity and mortality in people in care [73]. Thus, alternative approaches have been recommended both in respect to using a lower cut-off for treatment switch and switching individuals after the first high viremia [73].

6.1.6 Clinical Scores For Client Management And Triage In the Public Health Approach

Risk stratification and triage are central to clinical medicine across different disciplines. From the APGAR score to assess neonates to the Framingham Cardiovascular Risk Score to predict cardiovascular risk, clinical scores are often used to guide the choice of prophylaxis and therapeutic interventions [74] and tailor treatment to the risk profile of the individual client [75]. Many of these scores are based on mathematical algorithms that predict risk based on surrogate variables which are then presented as a paper-based nomogram or score chart or integrated into web-based calculators, apps or electronic patient record systems to inform decision making both on individual and population level [76]. Diagnostic risk scores can predict the probability of the presence of diseases in an at-risk person or differentiate between individuals more likely to benefit from particular treatments or interventions. They thus allow to better target available resources and promise more cost-effectiveness.

Several clinical risk scores have been developed in for HIV related patient management in Africa mainly in to identify HIV negative individuals with high risk of HIV acquisition that then could be targeted with tailored preventive messaging [77, 78] or to determine Co-infections such as Tuberculosis [79] but also to improve diagnosis of clinical failure on ARV [72] [80, 81]. So far, clinical scoring does not play a large role in the public health approach and has not been integrated into treatment guidelines on a national or international level. However, with increasing digitalization

of the health care sector in developing countries which also results in increasing availability of electronic patient management software in clinical and public health projects, new opportunities arise to integrate scores and utilize prediction algorithm to guide the patient flow through the clinic, optimizing clinic efficiency and enforcing treatment standards. Scores could also support tasks-shifting to less specialised HCW in the context of community-based differentiated care models, or integrated care models where HIV is treated in clinics specialised in care of pregnancy, TB or the general outpatient department, where scores integrated in decision support systems could assist patient triage, management and referral of clients not responding well to treatment.

6.2 Study Objectives And Rationale

If viral suppression as a central outcome measure of the “last 90” is understood as a result of interacting dynamics between virus, host and the socio-cultural context including the health system, research exploring viral suppression needs to investigate these dynamics both on individual and health system level.

In this thesis, such an investigation is conducted using the Walter Reed Southern Highlands HIV Care Program (WRSHCP) as an example. The WRSHCP is a large treatment program which uses the national health system to deliver HIV treatment and care in western Tanzania thus information gained from the program evaluation may inform national HIV treatment policy as much as other settings with comparable population and health system characteristics. Using the data of a cross-sectional study that included a retrospective data collection component, the current research aims to:

1) Describe the population benefitting from the WRSHCP and the facilities within which this population accesses care.
   This objective will focus on the descriptive presentation of the study population and the study sites.
2) Investigate the ability of the WRSHCP to achieve “the last 90” on the programmatic level using three different cut-offs commonly used as binary outcome measures of treatment success.
3) Explore if differences in virologic suppression can be observed associated with regions, health care levels or individual Care and Treatment Centres (CTCs) using different virological cut-offs.

Towards this aim, an analysis will also include comparisons adjusting for population differences through Propensity Score (PS) as detailed in the method section 8.9.2. As an alternative method to regression analysis preferable in the presence of multiple variables, adjustment by Propensity Score can balance the distribution of observed baseline covariates between the different exposure groups [82-84] and even allows a causal attribution of outcome to exposure [82, 85-88].

4) Explore treatment outcome in subpopulations defined by time on treatment and identify factors associated with virologic failure on the client level.
5) Develop and validate a predictive clinical score that can assist in identifying individuals in the population who are likely to fail ART > 400 copies/ml based on a theoretical framework that conceptualizes individual treatment outcome as the result of a multifactorial interplay between individual and community factors. This score is aimed to be used in the facility-based setting of the national ART program in Tanzania but could be adapted to other populations such as community-based ART delivery or national programs in other developing countries. It is envisioned as a screening tool for health care workers – particularly clinical officers, doctors and other personnel who request viral load tests in ART clinics - to identify individuals with increased risk of treatment failure. In order to identify viral replication early when additional interventions still promise viral re-suppression, HIV viremia above 400 copies/ml would be the preferable outcome to be predicted over the higher WHO definition of clinical failure above 1000 copies/ml.
6) Use Decision Curve Analysis to explore the benefit of the score developed in objective 5 in the context of the WRSHCP.

7 Endpoints

7.1 Primary Endpoint

To estimate the proportion of program participants on ART for at least 6 months who have achieved viral load suppression in this study defined as >400 copies /ml.

7.2 Secondary Endpoints

1) To describe viral load suppression in the cohort according to:

   a. WHO virologic failure criteria defined as >1000 copies/ml.
   b. The most conservative cut-off of >50 copies/ml.

2) To assess the impact of the site, region, health care level and ART stratum on program outcome.
3) To describe in separate analysis client factors associated with viral suppression
4) To develop a theoretical contextual framework that conceptualizes individual virologic treatment outcome as a balance of underlying dynamics supporting or hampering virologic suppression.
5) To develop and validate a clinical score based on the framework developed able to predict virologic suppression in the study population using the definition of >400 copies/ml as a virologic failure.
6) To explore the clinical utility of the clinical score in the public health approach using Decision Curve Analysis.

8 Methods

8.1 Study Design

The RV288 is a cross-sectional, observational cohort study with a two-stage sampling design. In the first stage, a sampling frame was generated from the WRSHCP programmatic data and 7 study sites were selected stratified by region and health care level as detailed in section 8.3.1. In the second step, a sample size of a total of 700 participants was recruited. Each site contributed 100 participants that had been stratified by time on ART and selected through Probability Proportionate to Size (PPS) sampling as detailed in section 8.3.2. Following an informed consent procedure as described in section 8.3.2, a single visit was conducted that included a blood draw, clinical assessment and a patient questionnaire as specified in section 8.5. With random sampling in both stages, the study design resulted in a self-weighted study sample considered representative of the WRSHCP population. Figure 7 presents a flowchart of the two-stage study design and recruitment of participants.

8.2 Setting

The study was part of a program evaluation titled “RV288 - A Virological Assessment of Patients on Antiretroviral Therapy in the US Military HIV Research

Program/President's Emergency Plan for AIDS Relief (PEPFAR) – Supported Programs in Africa". This multi-country program evaluation selected a representative sample of 22 ART clinics from six MHRP-supported PEPFAR programs in Kenya, Nigeria, Tanzania, and Uganda. Overall, the study population across the different countries targeted a randomly selected sample, stratified by time on ART (6-12 months; 13-24 months; >24 months), of approximately 2600 MHRP-supported PEPFAR program adult participants (either 325 or 700 participants from each of the six programs) enrolled in the MHRP-PEPFAR-supported antiretroviral treatment programs on first-line ART for at least 6 months who had had at least one follow-up ART clinic visit in the last 6 months.

The Tanzanian country program was evaluated through the RV288d protocol, which is the focus of this thesis. This protocol targeted 700 program participants at 7 sites stratified by region, time on ART and health care level. The study was implemented between 2013 and 2014. All sites were visited one month prior enrolment start to conduct the site assessment and clients were then enrolled over a period of two months. Sites were consecutively included by region, starting from Mbeya referral hospital and ending in Rukwa District Hospital. No follow-up was planned or conducted.

8.3 Participants

Participants were selected in a two-stage sampling design with the first stage selecting study sites and the second stage selecting participants.

8.3.1 Study Site Selection

Inclusion And Exclusion Criteria For Sites:

Sites were eligible for selection if they had:

- Offered ART services for at least 12 months.
- Had at least 100 patients on ART.
- Received support from WRSHCP.

Site Stratification

Accounting for Region and Health Care Level strata, sites were selected as follows:

1. Within the program catchment area, Mbeya Zonal Referral Hospital (MZRH) in Mbeya Region was the only tertiary hospital, hence was included to represent the referral health care level (sampling probability=1).
2. Four regional hospitals were in the project catchment area, one in Mbeya and Ruvuma respectively (sampling probability=1) and two in Rukwa (sampling probability =0.5). As the study aimed for regional representation, a regional hospital was included from each region.
3. On the district level, one hospital from each region was randomly selected, taking geography, partner type, length of time offering ART services, number of patients on ART and logistical considerations into account.

The resulting study sites considered representative for the PEPFAR sites were:

Mbeya Region:

Mbeya Zonal Referral Hospital	(MZRH)
Mbeya Regional Hospital in Mbeya	(MbRH)
Mbeya District Hospital in Tukuyu	(MbDH)

Ruvuma Region:

Ruvuma Regional Hospital in Songea	(RvRH)
Ruvuma District Hospital in Mbinga	(RvDH)

Rukwa Region

Rukwa Regional Hospital in Sumbawanga	(RuRH)
Rukwa District Hosptial in Mpanda	(RuDH)

8.3.2 Participant Selection

Inclusion Criteria

To be eligible for study participation, study participants had to be:

- Willing and able to provide informed consent.
- Enrolled in an MHRP-supported PEPFAR program.
- ≥18 years of age.
- On first-line ART for at least 6 months.

- Had attended at least one routine ART clinic visit in the past 6 months.
- Willing to be interviewed and to provide a blood specimen.
- Registered at the clinic under a patient ID on the Pre-recruitment List through the random selection process described below.

Exclusion Criteria

Study participants were excluded if they met any of the following criteria:

- On second-line ART.
- Mental or physical incapacity with the inability to provide informed consent.

Selection Procedure Of Study Participants

Four weeks prior study initiation at any of the study sites, the respective site was visited by the Principal Investigator and a complete list of patient identifiers and their respective ART strata was collected for all active subjects at this site. This list constituted the sampling frame for participant selection.

Clinic identification numbers were randomly selected from this list by an independent statistician and a Pre-Recruitment List was produced for each ART strata. To achieve Proportional Probability Sampling (PPS) by ART strata, a recruitment target was set for each stratum so that the proportion of study participants in this stratum reflected the proportion of clients in the patient population in this stratum at this site.

In a second step, the clinic identification numbers on the Pre-Recruitment Lists were matched to the corresponding patient medical records. If according to the medical record clients were likely to meet the inclusion criteria, they were approached for study participation. Sites with patient tracking procedures in place scheduled appointments for study enrolment along with the participant's routine clinic visits. Those sites without patient tracking procedures waited for patients to report for their routine ART clinic appointments. Each site recruited participants in the order as they appeared on the Pre-Recruitment List until all stratification groups were filled. If a client could not be enrolled, the reason for non-enrolment was documented in the Pre-Recruitment List. At each site, patients were recruited over 4-8 weeks. Figure 4 depicts the principle of the participant selection process, Figure 7 describes the recruitment outcome of the study population.

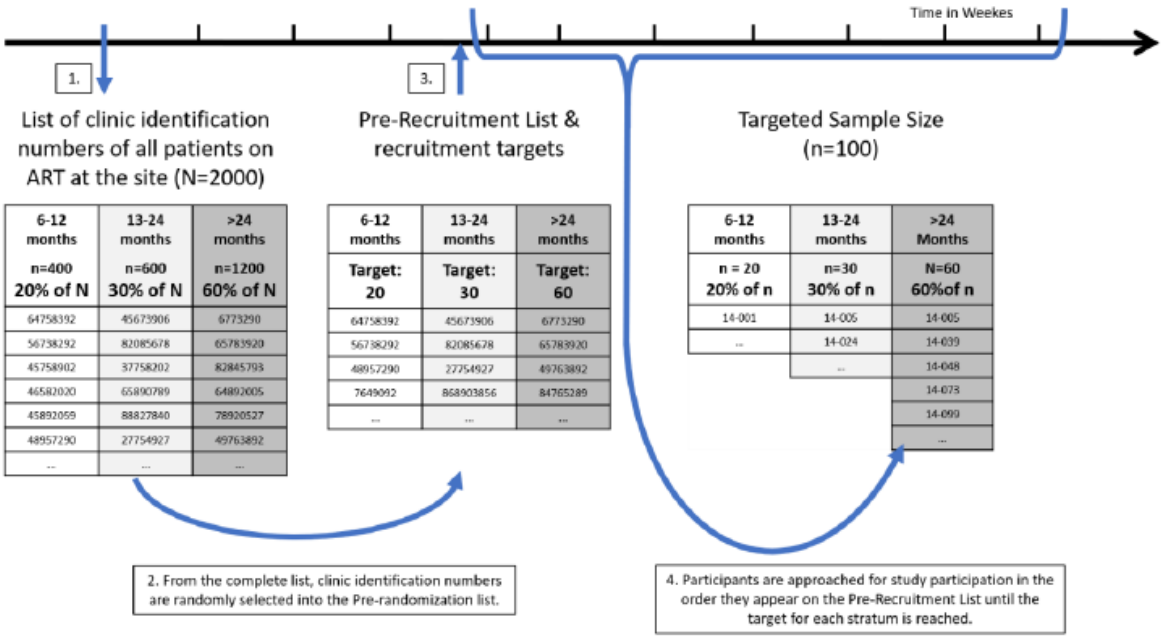


Figure 4 Principle Of Selection Procedure For Study Participants Using A List Sampling Frame
This figure demonstrates participant selection using a fictive site with 2000 participants on ART > 6 months with 20% in the 6-12, 30% in the 13-24 and 60% in the >24 month stratum: In the first step (1.) a complete list of clinic identification numbers by stratum is compiled. From this list, participants are randomly selected by a statistician independent from the study site, who prepares the Pre-Recruitment List and also provides recruitment targets for each stratum (2.). The list is returned to site (3.) that starts recruiting participants in order given by the Pre-Recruitment List (4.). The study population thus is representative of the study site population.

Informed Consent Procedure

Once clients reported to the clinic, they were approached for study participation. Written consent forms were available both in English and Kiswahili. Potential subjects had the consent form administered to them in a confidential environment by a trained member of the study team in the language best understood. For illiterate subjects, an impartial witness separate from the study staff was sought to witness the consent process following a study-specific Standard Operation Procedure that emphasized the voluntary participation and ability to withdrawal without penalty at any time.

8.4 Variables

The following definitions were applied for variables used in this analysis

8.4.1 Outcome Variable: Virologic Suppression

Virologic outcome was defined as binary endpoint Virologic Success/Failure based on the blood HIV viral load as measured at the study visit. In alignment with the three

different cut-offs used globally as discussed in section 6.1.3, three binary endpoint variables (VS400, VS1000, VS50) were created to describe treatment outcome for the primary and secondary endpoints. VS400 was the primary outcome parameter for which sample size and power had been calculated. Figure 5 presents the names, definitions of the different cut-offs used and the respective endpoint level.

Name	Virologic Failure	Virologic success	Endpoint*
VS400	>400 Copies/ml	=/<400 Copies/ml	Primary
VS1000	>1000 Copies/ml	=/<1000 Copies/ml	Secondary
VS50	>50 Copies/ml	=/<50 Copies/ml	Secondary

*Sample Size was powered for the primary endpoint only.

Figure 5: Definitions of Binary Outcome Variable "Virologic Success/Failure" Used In This Analysis.

8.4.2 Individual Level Exposures Assessed As Potential Confounders, Predictors and Effect Modifiers In The Analysis.

In line with the perception of virologic failure as a multifactorial outcome, additional information about the sites and individuals on treatment were collected as exposures with potential to confound, predict or modify the outcome depending on the method and analysis aim. Below, they are listed by thematic area:

Immunologic Parameters

CD4 counts at study visit, prior treatment start and available highest cd4 count were collected during the study and further immunological outcome parameters were derived from their relationship as follows:

Immunological failure (IF) was defined in line with the Tanzanian National Guidelines [89]. Based on available CD4 count development over time, IF was present if at least one of the following criteria was met: i) CD4 count at study visit dropped below baseline or ii) CD4 count dropped below 50% of peak CD4 value on treatment or iii) None of the available CD4 count on treatment was >100.

Immunologic Success (IS) was defined as an improvement in the immunological status characterized as an increase in CD4 count of at least 50 cells/mm3 per year or CD4 cell count >350cells/mm at study visit [2]. Where baseline or peak CD4 count was missing, cases were classified as unknown.

Clinical Parameters

Next to WHO stage defining events prior, during ART and on a study visit, specific non- communicable diseases of interest were collected. These included non-communicable diseases (Hypertension, Diabetes, heart, renal and liver disease) (ii) common co-morbidities of HIV infection (Anemia, Cancer, Tuberculosis, sexually transmitted infections) and diseases in the neurologic-mental health domain (migraine, seizures and depression) (diabetes, cardiac events, liver diseases). Additionally, a detailed assessment of symptoms observed within three months prior the study visit included fever, skin rash, mouth sores, sore throat, shortness of breath, unintentional weight loss, night sweats and cough, fatigue, muscle and joint aches, loss of appetite, nausea, vomiting, abdominal pain, diarrhoea, and neurological symptoms such as confusion, headache and stiff or painful neck.
All disease events and symptoms were assessed if they met the WHO clinical case definitions [90] and classified as either non-WHO or WHO stage 1 to 4. Clinical failure was defined as new or recurrent WHO stage 4 disease 6 months after treatment start [91]. Immune Reconstitution Syndrome (IRS) IRS was met if a stage 4 WHO defining disease was recorded within 180 days after initiation of ART. Clinical success was defined as an improvement of WHO T-stage - which was the WHO staging at study visit without considering previous WHO disease – against the WHO stage at treatment start. The study further collected HIV disease-specific information such as time of HIV diagnosis, weight and functional status at treatment start and vital signs (weight, height, temperature, Blood Pressure and Respiration rate) at the study visit.

Laboratory Values:

The following laboratory values were collected:
Closest to the time of treatment start (retrospective data extraction): CD4 count, Haemoglobin, Lymphocytes, Alanine aminotransferase (ALT), Creatinine, Wight and Body Mass Index prior ART. On Study visit: CD4 count, Red Blood Count, White Blood Count, Haemoglobin, Haemoglobin %, MCV, Neutrophile, Platelets, Monocytes, Monocytes %, Lymphocytes, Lymphocytes %, Eosinophile, Eosinophile %, Basophile, Basophile %, ESR, Creatinine, ALT.

Existing Antiretroviral Therapy and Co-medications:

The components of the ART regimens used were collected separately and two groups were formed identifying clients by their Non-Nucleoside Reverse

Transcriptase Inhibitor (NNRTI) use and the Nucleoside Reverse Transcriptase Inhibitor (NRTI) backbone of their regimen, ever received prevention of mother to child prophylaxis, use of Pneumocystis Jiroveci Pneumonia (PCP) and TB prophylaxis.

Study Design Variables:

Region, Health Care Level, ART Stratum

Demographic Variables:

Age, gender, literacy, highest formal schooling obtained,

Socio-Economic Status:

marital status, occupation, number of financial dependents, adults and children living in the household and availability of electricity in the household. Costs, time and mode of clinic access for the individual.

Adherence And Health-Related Behaviour

The number of missed clinic visits, participation in support groups and having a treatment supporter, use of traditional healers and remedies in parallel to ART. Alcohol consumption per week, smoking and drinking habits.

Interaction With The Health System:

Duration of treatment cascade from diagnosis to ART initiation as a whole and each individual step (diagnosis-enrolment in care-eligibility for ART-ART initiation).

8.4.3 Site Level Exposures

The following Site Level variables were assessed in the context of this analysis:

CTC specific information included: type of adherence counselling (group, individual or combined), number of adherence sessions prior ART initiation, tracing of patients lost to follow up and weeks after which such tracing is initiated, year of ART program initiation, patients ever and currently enrolled on ART, including total number of patients on ART disaggregated by first or second-line regimens. Inter-site referral of critically ill and stable patients. Number and cadres of Health Care Workers (HCW) assigned to the CTC. Further, the site level exposure variables included population size of the catchment area as per 2012 national household survey and population size of the town in which the facility is situated. From this information, the following

variables were derived: health care worker to patient ratio, number of patients lost per patient retained, the average number of patients lost per year of operation and as the percentage of the current patient population and the ratio of the total population on ART per clinic day.

8.5 Data Sources And Assessment

Data was collected through a site visit assessment and during the study visit that included a patient assessment, retrospective data extraction and collection and analysis of blood specimen. Information was collected in source documents and then transferred into four paper-based collection tools provided as templates in Appendix 13.5:

1) A Site Observation Questionnaire
2) A Nurse Administered Questionnaire
3) A Participant Specific Case Report Form

All steps of the study procedures were defined by the study protocol and a manual of procedures and performed by specifically trained study staff. Data collection procedures are summarized in the following.

8.5.1 Site Level Assessment

A structured site assessment was conducted within 4 weeks prior to recruitment of participants at study sites. During these site assessments, the Principal Investigator of the study discussed with a team of health care workers (HCW) following the assessment in the site observation questionnaire. Following this discussion, a site visit was performed which collected information on the availability of resources in a structured fashion. The Site Observation Questionnaire is provided in Appendix 13.5.

8.5.2 Nurse Administered Patient Questionnaire

A trained health care worker administered the patient questionnaire in a confidential setting. Translations for the questionnaire were available in English and Kiswahili and the interview was administered in the language best spoken by the client which was Kiswahili in all times. Participants were encouraged to answer to the best of their knowledge but always had the option to decline a response. The information was collected separately in the Nurse Administered Questionnaire (Appendix 13.5).

8.5.3 Clinical Assessment And Medical History:

Through patient interview and review of the clinical file, a general medical history was taken by a clinical officer trained in study procedures. The clinical officer further recorded the presence or absence of clinical symptoms for the period of 3 months prior study visit and at study visit through triangulation of patient self-report, documentation in the patient file and clinical observation. A complete physical examination was performed at study visit, and abnormalities by organ system were documented as free text description, collecting both diagnoses and symptoms. Vital signs were collected (body weight in kilogram), the temperature in degrees centigrade, pulse in beats per minute, respiration in breaths per minute and blood pressure following Standard Operation Procedures using calibrated equipment. The clinical officer then assessed all diseases entities comprised in the WHO clinical case definition [90] separately in respect to lifetime presence, onset in relation to ART start and if present at the study visit.

Following the completion of the study visit, a trained health care worker completed the medical record abstraction form that was part of the Participant Specific Case Report From extracting from the clinical patient file.

Blood was collected at study visit and laboratory analysis was performed at the local and central study laboratories.

The laboratory of the study site performed the following assessments:

- CD4 count (Facs Calibur from Becton Dickinson),
- Haematology (XT 1800i Sysmex) and
- Erythrocyte Sedimentation Rate (ESR, Westergreen method)
- Plasma separation for sample transportation

The Mbeya Zonal Referral Hospital laboratory or its backup-laboratory at the Mbeya Medical Research Center performed the following assessments:

- Clinical Chemistry (Cobas INTEGRA 400 plus (Roche)) and
- HIV-PCR (Cobas TaqMan 48 HIV-1 test version 1.5 (Roche Diagnostics, NJ, USA).

Samples were shipped maintaining a cool chain of -20 degrees from site to the central laboratory.

8.5.4 Comparability Of Assessment Methods

Quality Assurance Of Clinical Data Collection

Prior to the study initiation at each site, study staff was trained in a one-day study-specific training that included all study-specific procedures relevant to their specific roles. Following the completion of the Case Report Forms, the forms were reviewed for completeness and validity by a second study team member on-site before they were transferred for data entry.

Quality Assurance Of Laboratories

Standard operation procedures, tests of laboratory competency, documentation of quality assurance measures and inter-laboratory validation runs with the College of American Pathologists accredited laboratory for all parameters ensured laboratory comparability. All laboratories involved in the study underwent targeted training prior to the study initiation and were required to meet defined quality standards for the respective test prior to the study initiation at that site. Personnel performing laboratory tests were blinded to clinical information. The central HIV laboratories were located at the Mbeya Zonal Referral Hospital (MZRH) and supported by the laboratory of the National Institute of Medical Research Mbeya Medical Research Centre. The latter is accredited with the College of American Pathology, providing a wide panel of haematology, clinical chemistry, immunophenotyping, HIV-RNA (COBAS TaqMan HIV-1), and immunological assays.

8.6 Bias

The study design aimed to ensure a study population representative to the overall program population. However, as a cross-sectional study, results of RV 288 are limited in the following respect:

- Selection bias: In RV 288, the following patient group had a lower probability of recruitment: Patients enrolled in the PEPFAR Program but lost to follow up or died prior to study enrolment, HIV infected people in need of treatment and living in the catchment area but not accessing treatment in the PEPFAR program, patients on treatment at the respective sites but not captured in the database or treatment register that is used to generate the randomization list, patients approached for the study but not willing to give informed consent. In the face of a high number of undiagnosed HIV infections in Tanzania, a substantial amount of

HIV infected persons in the regions of interest were not represented in the study sample. The study results, in consequence, were biased positively and caution should, therefore, be applied when generalizing to all patients ever started in the PEPFAR program or to the target population of HIV infected people in the Southern Highlands at country level, and HIV infected population in the catchment area of PEPFAR programs in respective countries on the program level.

- Information bias was especially relevant in the Site Observation Questionnaire applied to all sites and aiming to assess the site and quality of services provided and the Patient Questionnaire aiming to asses adherence. In both cases, the number of interviewers was limited and emphasis was put on specific training for study personnel responsible for assessing those aspects.

To counteract selection bias, we applied the following measures:

- Stratification for the applicable program, region and level of care when selecting the sites involved in the study and stratification of the patients according to time on ART on subject recruitment level.
- Assessment of a number of patients not enrolled for any reason through the screening log.
- Review of status of databases and treatment registers during pre-site initiation visits to all sites involved in the study.

8.7 Study Size And Sample Size Calculation

8.7.1 Sample Size Calculations For Data Collection And Determination of Primary Endpoint

The sample size calculation was based on an expected refusal rate of 0% at the site level, <5% at the patient level and 12% virologic failure after 12 months.
For point estimates for the primary endpoint – virologic suppression on the program level - 700 patients were needed for estimates between 8% - 50% (0.05 precision and 95% confidence interval). This sample size further allows point estimates between 8% and 30% for secondary endpoints. To compare different health facilities with 80% power and an alpha of 0.5, a total of 98 participants had to be recruited from each site.
Additional sample size calculations were performed to evaluate virologic suppression on health care level and across regions.

8.7.2 Sample Size Considerations For The Development Of The Predictive Score

As the sample size was determined by the primary endpoint of the study, which was virologic outcome above 400 copies/ml, there was no specific sample size calculation for the derivation of the score. For predictive models, the number of events to be predicted is more important than the number of participants in the sample and at least 10-20 outcome events per variable is considered acceptable [92, 93]. In our study, we had 458 participants in the training dataset, of which 75 had virologic failure above 400 copies/ml. We hence were planning to include not more than 7 variables into the predictive model.

8.8 Quantitative Variables

The following presents the main conversions of quantitative variables during analysis.

8.8.1 Outcome Variables:

Virologic Suppression And Virologic Failure

Based on the considerations presented in section 6.1.3, we used the continuous variable of viral load at study visit to group participants into three binary endpoints of treatment success or failure defined as follows:

VS400: Virological failure defined as the viral load at study visit above 400 copies/ml,

Virological suppression defined as the viral load at study visit equal or below 400 copies/ml.

This definition was the primary endpoint of our study, on which power calculation was based.

VS1000: Virological failure defined as the viral load at study visit above 1000 copies/ml, Virological suppression defined as viral load at study visit < or equal to 1000 copies/ml.

This definition was a secondary endpoint using the cut-off favoured by WHO

VS50: Virological failure defined as the viral load at study visit >50 copies/ml, Virological suppression defined as viral load at study visit below or equal to 50 copies/ml.

This definition was a secondary endpoint using the cut-off commonly used in resource-rich countries.

8.8.2 Confounder Variables:

Body Mass Index (BMI)

The BMI is a measure for indicating nutritional status in adults defined as a person's weight in kilograms divided by the square of the person's height in metres (kg/m2). The BMI at study visit and at treatment start were calculated using the weight recorded as outlined above. Results were then grouped using a simplified WHO classification: <18.5 underweight, 18.5–24.9 normal weight, >24.9: obesity [85] [94].

Reclassification Of Variables With Low Cell Values

To avoid low single values in some of the categorical variables across regions, variables were collapsed prior analysis:

- The variable accessing the use of support groups combined the answer possibility "I don't know" and "No".
- The number of missed refill visits were collapsed to a binary measure of has/has not missed any re-fill visit.
- To account for missing values in the baseline safety laboratory assessment (Lymphocytes, Haemoglobin, Creatinine, ALT), variables categorized as described above were collapsed into 3 categories (normal, abnormal, missing) using regional normal ranges [95].
- Functional status was combined with binary (working, others)
- Drug regimen were collapsed to PMTCT binary "yes, any" or "no"

8.9 Statistical Methods

Data were analysed using Stata Statistical Software Package version 12 or 16 [96]. Which version was used is stated in the respective method section.

8.9.1 Descriptive Statistics

Summary statistics for all outcome variables and confounders are presented for the total study population and by site with missing values stated.

Continuous variables include the mean and standard deviation, categorical variables include frequency and percentages. WHO relevant disease events are presented disaggregated by stage and individual diseases and in relation to ART initiation.

For single variable comparison of unadjusted data, Kruskal –Wallis rank test was used for comparisons of medians and Pearson Chi-square test was used for comparison of proportions.

8.9.2 Virologic Outcome Analysis

In line with the objectives of this research, virologic suppression was assessed at three different virologic cut-offs relevant in the context of international public health decision making as detailed in section 6.1.3., using three binary virologic cut-offs (VS50 VS400 and VS1000) derived from the continuous variable viral load at study visit as described in section 8.8.1. A 95% confidence interval was generated around the point estimates.

In line with STROBE Guidelines and the aim of this study, analysis is presented with different levels of adjustments:

(i) "Outcome in the Study Population" –presents the unadjusted outcome in the study participants as discussed in section 8.11.2

(ii) "Outcome in the WRSHCP Population" –presents the estimates that apply for the WRSHCP Population of which the Study population was only a sample. Here, study design is adjusted for as discussed in section 8.11.1 and the point estimate and confidence interval are estimates applicable to the full Walter Reed Southern Highland Care and Treatment Program (WRSHCP).

(iii) "PS Adjusted Population" present the results of the sub-group comparisons controlling for pre-treatment differences through Propensity Score (PS) weighting as described below.

<u>Virologic Outcome Controlled For Baseline Variables Through Propensity Score Methods</u>

We assessed the outcome of sub-groups defined by site, health care level and the region controlling for differences in the study population through Propensity Score Methods.

The Propensity Score (PS) as defined by Rosenbaum and Rubin [87] is the probability of a study participant to be in a specific exposure group (in case of the current analysis the study site, region or health care level) as derived from the baseline characteristics of the study population. Adjustment of the study population by the Propensity Score through stratification, matching or weighting [82-84] can

balance the distribution of observed baseline covariates between the different exposure groups. The average effect of the exposure on the outcome can then be estimated in this PS adjusted sample, as the balanced sample allows the estimation of the counterfactual outcome for exposed subjects from the control subjects [82-84] given the assumptions of “overlap” and “strong ignorability” hold:

“Overlap” describes a situation where the probability of a study participant to be in a specific exposure group is non-zero for more than one exposure group. Overlap thus implies that there are no values of variables that occur only at one of the exposure groups. Overlap can be assessed in the data and the respective overlap plots for the analysis conducted in this thesis have been included in the supplemental material provided in section 13.4.

The assumption of “strong ignorability” as the second assumption of PS methods states that next to the variables used to derive the Propensity Scores, there are no further unknown relevant confounders that influence the outcome of interest. “strong ignorability” by definition cannot be ascertained as it concerns unknown confounders [97].

If these two assumptions are met, Austin states that a PS method “... mimics some of the particular characteristics of a randomized controlled trial” [82], as any difference between outcome in the exposed and unexposed group in the presence of the PS can be causally attributed to the exposure” [85-88].

Compared to regression analysis as an alternative method to adjust for confounders, PS methods are recognized for several advantages: They can reduce multiple confounders into a single score and thus avoid the multi-dimensionality that often restricts the number of variables that can be included in a regression model [98-100]. Further, contrary to regression methods that combine outcome analysis and control for confounders in one model, PS methods balance baseline characteristics through an independent step that is separated from the outcome assessment. Through this separation, bias is prevented as the regression model selection process cannot be influenced by desired outcomes [101].

Especially in datasets with few events per confounder, Propensity Score Methods have shown to be less biased and produce more robust and precise estimates than logistic regression models In simulation studies [98], making PS methods the favourable method for the outcome analysis controlling for population baseline differences in the RV288d dataset.

We derived Propensity Scores (PS) using the Generalized Boosted Model Technique (GBM) that is described in detail below. To compute the score, we followed the procedure developed by Cefalu et al. [102] which uses Stata Macros developed by the RAND Corporation to create an interface through which stata can access the twang package in R [103] [102, 104]. Through this interface, the mnps command in twang – which has been designed to develop PS for exposures with more than two groups – can be used by stata [102, 103] and the thus derived PS scores can be fed into the outcome analysis.

Accordingly, we generated PS scores and in a second step, we applied a PS weighted logistic regression model to assess the impact of the exposure on the three binary outcome measures (VS 50, VS400 and VS1000). The “double robust method”, which adjust for PS and in addition includes variables, where a full balance has not been achieved through the PS in the regression model, has shown to provide the most balanced baseline variable estimates in simulation studies [85, 97, 105, 106]. We hence applied the “double robust method”, included selected variables where full balance had not been achieved with the PS score weight alone. For weighting, we used the syv commands in stata [107] as described in section 8.11, replacing the probability weights for the Propensity Score. This allowed us to further account for the study design.

Pairwise outcome measures commonly referred to as the Average Treatment Effect (ATE) were used to describe the impact of the respective exposure. ATE is a summary measure that compares estimated mean outcomes if the entire population been observed under one exposure versus had the entire population been observed under another exposure [97].

We now discuss in detail the considerations that guided variable selection, describe in more detail the GBM method and the assessment of balance and overlap.

Variable Selection

Guided by existing literature [82, 88, 97, 108], we considered all potential confounders as presented in section 8.4.2. and identified demographic and baseline variables. Variables that are only related to the exposure but not to the outcome have been shown to decrease precision without proving balance in simulations studies evaluating binary Propensity Score Methods [108], so only known or assumed confounders to treatment outcome were selected for further assessment.

Following the procedure described by Spreeuwenberg et al [88], we selected all variables that were associated with the VS50, VS400 or VS1000 outcome with a significant level of $p<0.4$ in univariate logistic regression analysis. We further included time on ART and years with HIV infection at study visit to account for the retrospective nature of the dataset. As the Generalized Boosted Model Technique (GBM) automatically includes indicators for missing values in the model [97], we included the continuous variables even if they had several missing values rather than categorical variables that were applied in the logistic regression for patient-specific factors as discussed in section 8.4.2.
The following variables were included in the Generalized Boosted Model as described below :
Individual characteristics: gender, age, education, profession, drinking and smoking habits,

Relationship to others: Marital status, number of financial dependents, number of adults in the household and access to electricity.

Relationship to the clinic: Mode of transport to the clinic, and distance to the clinic, time on ART at study visit and years of HIV infection

Clinical presentation at ART start: Body Mass Index, weight, Cd4 count, ALT, Lymphocyte count, Haemoglobin, history of the following WHO staging relevant diseases at baseline: Herpes Zoster (WHO stage2), Prolonged fever (WHO stage3), Tuberculosis (WHO stage 3 or 4), Loss of body weight (WHO stage 2 or 3).

Estimation Of The Propensity Scores Through The Generalized Boosted Model Technique (GBM)

The Generalized Boosted Model Technique (GBM) is a computational tool based on machine learning techniques which creates a constant model out of an iterative combination of regression trees to predict a binary outcome. McCaffrey describes the computational process as follows:
“The model consists of many simple regression trees [109] iteratively combined to create an overall piecewise constant function. The iterative fitting algorithm begins with a single simple regression tree and at each new iteration, another tree is added. The new tree is chosen to provide the best fit to the residuals of the model from the previous iteration. […] When combining trees, the predictions from each tree are shrunken by a scalar less than one to improve the smoothness of the resulting

piecewise-constant model and the overall fit. Each iteration increases the likelihood and with enough iterations, the model is sufficiently flexible to perfectly fit the data."[97] Pre-defined stopping rules determine which model is at the end selected to produce the Propensity Scores. These stopping rules aim are commonly based on either the Absolute Standardized Mean Difference (SMD) or the Kolmogorov Smirnov Statistic (KS) that compare the mean and distribution of the unadjusted and PS adjusted variables avoid model overfit [97] and are defined below.
Compared to other methods to determine the PS, GBS automates the iterative refinement process and has been shown to provide PS that allow a better balance between the different exposure groups. GBM was developed for the comparison of binary exposures but has also been applied successfully to generate PS for multiple exposure groups [97, 102]. The twang package in R - which was used in this analysis- generates PS scores for multiple exposures through a series of binary comparisons, each comparing a selected group of the exposure with the cumulative data of all groups not in this group.

GBM - assessment of balance and overlap

For each exposure category, we developed a GBM model using the variables selected as described above. Balance of adjusted variables and population overlap were assessed through a series of visual plots centred around the stopping rules used in the GBM model (Absolute Standardized Mean Difference (SMD) and the Kolmogorov Smirnov Statistic (KS)) following the procedure delineated by Cefaru [102]: First, optimization plots of the respective stopping rule (SMD or KS) against the number of iterations were used to ascertain that sufficient iterations of the model trees were run and thus the best model selected. Overlap was then assessed using visual inspections of the distribution of the Propensity Scores across the sites and is included as supplementary material presented in 13.4.1.

The main measure used to assess balance was the Absolute Standardized Mean Difference, which for our outcome of interest - the ATE - in a sample with more than two exposures is defined by McCaffrey as the "absolute value of the difference between the mean for treatment group and the mean for the control group divided by unweighted standard deviation of the pooled sample"[97]. Generally, standardized mean differences of less than 0.25 are considered small, 0.40 are considered moderate, and 0.60 are considered large [110]. Variables with the mean SMD above

0.2 in the weighted sample were included as additional variables in the double robust logistic regression.

The Kolmogorov Smirnov Statistic (KS) was used to assess the distribution of the differences between weighted and unweighted variables and compare different models. In line with McCaffrey, we considered KS statistics greater than 0.1 as signs of imbalance for the comparison of Region and Health Care Level but not for the comparison of sites with their much smaller sample size [97].

Next to assessing individual variables, we used the overall summary measure Suggested by McCaffrey of taking the maximum of the balance metrics for each treatment [97].

8.9.3 Additional Analysis - Influence Of Patient-Level Characteristics On Treatment Outcome

To identify factors associated with treatment outcome, we first assessed the distribution of those variables against the respective binary outcome of virologic suppression below 1000, 400 and 50 copies/ml (VS1000, VS400, VS50) using Kruskal-Wallis test for comparisons of continuous variables and Chi-square test for comparison of proportions and a significant threshold of $p<0.2$.

All variables thus identified were ordered by domains and the order of these domains were fixed across the models using the VS400, VS1000 and VS50 cut-off. Domains were as follows:

1) Non-modifiable variables such as ART stratum, region, health care level, or PMTCT history or pill burden
2) Clinical information such as clinical failure or specific diseases.
3) Laboratory variables of the study visit.
4) Laboratory variables prior to study visits, such as peak or baseline CD4cells
5) Adherence parameter and socio-economic contexts such as reasons why drug intake might be missed and patient satisfaction of services

All variables were then included in a logistic regression model which reported Odds Ratio (OR), respective confidence interval and Fisher's exact p-value was created in a stepwise backward selection process for each of the virologic outcomes. Akaike Information Criterion was used to assess model fit. All reported p-values were two-

sided and for all statistical tests an alpha level of <0.05 was used to define significance.

8.9.4 Development Of A Clinical Score To Predict Virologic Failure Above 400 copies/ml.

A further objective of this research was to develop a diagnostic clinical score to predict individual treatment failure in adult patients receiving HIV treatment in Tanzania through the PEPFAR supported national treatment program as outlined in section 6.2. The score was designed to predict HIV viraemia above 400 copies/ml, but its performance to predict clinical failure above 1000 copies/ml were to be additionally explored in a separate analysis.

To develop the score, we first split the dataset into training and validation datasets and then developed a theoretical framework for treatment failure and a multilevel diagnostic model based on this framework using the training dataset. This model was validated internally through bootstrapping and externally using the validation datasets. Throughout, the methods applied aligned with the approaches proposed by Steyerberg and Vergouwe [76] and Lee [111] and are reported guided by the TRIPOD Statement [92, 112] for quality reporting of clinical predictive models. In the current section, we focus on the description of the statistical analysis methods in respect to model development, validation and transformation to nomograms.

Model Development

Generating The Training And Validation Dataset

To generate training and testing datasets, we used the stata “random” function to assign random numbers to the individual cases and then rank the cases by these random numbers from low to high. The first 150 cases were sett aside as validation dataset. We further selected the complete population of the treatment site of the case with the lowest random number as the testing site so that we could evaluate the predictive capacity of the model for a patient population with an unknown site.

Development Of A Theoretical Framework

Prior to variable selection, a theoretical framework developed based on existing literature reviews that conceptualize individual treatment outcome as the result of a multifactorial interplay between individual and community factors, with the health care

system being considered a specially relevant area of the community. These dynamics were assumed to cumulate in two causal pathways that describe viraemia as the variable through which virologic failure or success is defined. The framework is presented in section 10.2.1. and formed the basis for variable selection.

Variable Selection

For a variable to be considered for inclusion in the predictive model, it had to be considered a meaningful surrogate parameter of the underlying dynamics described in the theoretical framework. Thus in a first step, all variables collected in the study were assessed in the context of the theoretical framework and either paired with the respective step of the causal pathway for which they were considered meaningful surrogate parameter or excluded from further consideration. Paired variables were then assigned a position according to the step in the causal pathway they represented. The closer to the outcome measure viremia (considered more “downstream” in the causal pathway) a step was, the better the position associated with this step and the respective paired variables.

In a second step, all paired variables were ranked in comparison to other variables that had the same position as they had paired with the same step of the causal pathway. Aspects considered in this ranking were potential competing physiological dynamics that would impact the respective surrogate parameter independently of the causal pathway, number of events in the study population, association of the variable with the outcome in univariate analysis and evidence of the association between the variable and viral load suppression from existing peer-reviewed literature. Ranking of variables was further influenced by the intended user and setting in mind: As we envisioned a trained health care staff delivering HIV treatment and care in a clinical setting - most likely doctors or clinical officers who would also make the decision on viral load testing - This would allow the inclusion of variables that require clinical knowledge and the ability to make diagnostic classifications such as to correctly apply the WHO staging, verify the presence of clinical symptoms, the correct identification and differentiation between different classes of drugs or drugs used for prophylaxis and treatment. Further, the clinical setting would allow to include laboratory variables as much as basic clinical examinations such as height, weight, pulse or temperature, for which basic medical equipment would be needed. However, keeping stockouts and laboratory shortages but also potential use in community-

based settings in mind, laboratory diagnostics available as point-of-care or “bedside” tests were prioritized over tests that require laboratory equipment.
The ranks were assigned through expert opinion triangulating existing information and considerations as follows:

Rank 1: Variable shows an overall preferable profile for model inclusion

Rank 2: Variable is meaningful but less so than other variables paired to the same step in the causal pathway.

Rank 3: Other reasons exist to not include the variable as per expert assessment, such as variable is correlated with treatment outcome in literature or but not in the study sample, it does not promise to add additional value to the model next to other variables with higher ranking paired to the same step in the causal pathway, can be expected to become irrelevant due to future changes in treatment guidelines or has very little variance in the sample population.

Rank 4: Substantial reasons exist that prohibit variable inclusion such as a very low number of cases or significant missing values in the training dataset

Due to the low variance in the site-level variables (Level2), variable inclusion was highly limited by co-linearity. In these variables, we additionally assessed correlation matrices and ranked variables by the amount of correlation with other variables within their group. Variables were then ordered first by their rank compared and then by the position of the step in the causal pathway they represented. This order defined their entry into the prediction model and is in detail presented in the supporting documents 13.4.2.

Model Estimation

Considering the study design as detailed in 8.11.3, we aimed to develop a diagnostic multilevel mixed logistic regression model that included participant-specific variables as fixed Level1 effects (individual characteristics) and Level 2 effects (site-specific characteristics) as much as a random intercept for the sites. The model was estimated using the “melogit” command in stata 16 [113], which fits mixed-effects logistic regression models for binary responses using the Bernoulli distribution to model random effects.
In the initial complete case analysis, the model was fitted to all cases with complete data in the training dataset. Starting with an empty model to determine, a full model was defined which included the eleven highest ranked individual-level parameters

which represented variables considered most suitable for model inclusion and most downstream steps in the causal pathways. This model was reduced through a backward selection process guided by AIC and BIC.
Continuous variables that were then one-by-one re-scaled and centred if appropriate, guided by visual inspection of their distribution plots, AIC and BIC to assess the impact of different scales on model fit. Exclusion of the variables that had been removed in the first step was then confirmed by entering them again one by one into the model containing scaled and centred continuous variables. Next, groups of variables paired to pathway steps further upstream were assessed for variables that could improve model-fit. A group-wise forward-backwards selection process was used to explore different combinations that were considered clinically meaningful. Interaction terms between the variables remaining in the model were explored. In this respect, we specifically considered the possible influence of age to either attenuate or aggravate the impact of other exposures, interactions between measurement and context of measurement, between calendar time and treatment and interactions between quality and quantity of risk factors [114]. However, only a few meaningful interactions could be identified as theoretically impacting viral suppression and none of them showed a clear benefit to model fit. As interaction terms rarely add to the predictive ability of a model and selective interaction term inclusion may drive model overfitting [92], interaction terms eventually were not included.
Finally, fixed site-level (Level2) effects were included in the model with first entering variables that highly correlated with other site-level variables followed by site-level variables that were not correlated with Level 2 variables maintained in the model. In this step, fixed site-level variables that reduced the random effect were favoured next to those improving AIC and BIC with the aim to generate a model that could support prediction in sites not represented in the training sample. Inclusion of random slopes into the model fit did not improve AIC or BIC.

Assessment Of Model Performance

Overall performance is the mathematical distance between predictions and outcomes [115] and was determined for the model both in the training (Apparent Performance) as much as in the external validation dataset. To assess performance, we used the following parameters:

Calibration

Calibration describes the correctness of prediction and extent of bias of a model by comparing predicted and observed endpoints [76, 115, 116]. We assessed calibration through visual plots of the Hosmer-Lemeshow test that compare observed and predicted probabilities. Smoothed lines overlaid over the plots ideally should have an intercept of 0 and a slope of 1 for best prediction [76]. Perfect prediction of the outcome from a 45-degree line in this calibration plot. In the validation dataset, a higher or lower intercept indicates systematic bias and resulting continuously too high or too low-risk predictions (“calibration at large”), while a slope lower than 1 may be the result of regression to the mean or indicate overfitting of the model to the training dataset. Better calibration is achieved if the Hosmer-Lemeshow test p-value is large, indicating no difference between the predicted and observed outcomes across the quantiles.

Discrimination

Discrimination describes the ability of the model to correctly differentiate between patients with the endpoint and patients without [76] and is commonly assessed through the concordance statistics (C-statistics) [115, 116]. For a binary endpoint, the area under the receiver operating characteristic curve (ROC-AUC) is the graphical equivalent of the C-statistics, when the true positive rate (sensitivity) is plotted against one minus the false positive rate (specificity) [76]. The larger this area under the curve, the higher is the discriminatory properties of the model. A ROC-AUC of 1 represents perfect prediction, a ROC-AUC of 0.5 represents chance [92]. As ROC-AUC - opposite to the positive and negative predictive value as alternative measures of discrimination - do not depend on event prevalence, they are preferred for predictive models that should work in settings with various and changing prevalence [117].

Model Validation

Internal Validation

For internal validation, we describe model calibration and discrimination (apparent performance) [118] in the sample dataset within which the model was developed (the training dataset). To quantify the model optimism of the model as a measure for model overfit in the training database, we followed the TRIPOD recommended approach as follows[92]:

1) A bootstrap sample the same size than the training dataset was generated from the testing dataset by sampling 200 individuals with replacement from the original sample. As simulation studies in clustered data demonstrate that accurate estimates of optimism can be obtained from re-sampling either on a cluster or participant level, but estimates are overestimated when a sequential sampling strategy from both levels is applied, we drew the bootstrap samples on patient-level only [119].
2) The multilevel logistic model was fitted to this bootstrap sample. As initial variable selection had been theory-driven and further model modifications during the model fitting process had balanced AIC and BIC improvement against model complexity and expected field practicability, the variable selection process as a whole was not repeated in the bootstrap sample.
3) The apparent performance (C-index for performance at large, Hosmer-Lemeshow test for calibration and for discrimination) of the model in the bootstrap sample was determined. (Bootsrap performance)
4) The performance of this model in the dataset not included in the bootstrap model was determined. (Test Performance)
5) The optimism was calculated as the difference between the bootstrap performance and the test performance.
6) Steps 1 to 5 were repeated 150 times and an average was generated from the estimated optimisms in step 5 that then represented the estimated optimism of the model in the test database.
7) This estimated optimism was then subtracted from the apparent performance to obtain an optimism-corrected estimate of performance.

External Validation

External validation provides information on model performance in populations unrelated to the training dataset [115]. As the strength of an external validation grows with the difference between the training and validation dataset, the random split as used in this thesis is often considered internal rather than external validation [76, 92] Nevertheless, as the score was aimed primarily as a programmatic tool in the Program Population population from which the study participants were sampled, the two validation datasets were considered sufficient for the primary use of the score. We had two overlapping datasets as testing datasets: One consisted out of the total population of one study site that had not been included in the training dataset. This

dataset was used to evaluate score performance in a separate location, providing a geographically separate testing dataset. The second dataset was the one which had been randomly chosen and included subjects managed at the sites that had been included in the training dataset and subjects that had been selected from the validation site.

8.9.5 Decision Curve Analysis

Decision Curve Analysis is increasingly used to evaluate predictive models in the context of their usefulness in clinical practice and is recommended as additional analysis in the TRIPOD guidelines as it complements mathematical measures of performance [92]. Decision Curve Analysis allows identifying the best strategy through which the benefit of an intervention can be maximized in a real-world setting. As it is a method that compares the benefit of an intervention relative to the benefit of other possible options, it does not require additional information but can generate a recommendation using only the dataset the new intervention was derived from.

Decision Curve Analysis recognizes that the importance of a false positive or false negative prediction derived from a model or test result can differ depending on its clinical use: In a case where a false positive prediction leads to an unnecessary intervention that carries significant risk – for example, pre-emptive mammectomy to prevent potential breast cancer - the trade-off between the expected benefit and the harm associated with the intervention will be different from a situation where a false positive prediction would have less drastic consequences - for example a vaccination to prevent Cervical Carcinoma [120]. Such trade-offs between harm and benefit of an intervention are continuously made in clinical or public health practice impacted not only by clinical but also by economical and feasibility considerations. While the choice of “treat” or “don’t treat” are relatively self-explanatory at the ends of the scale – if the risk for cancer is high, mammectomy is reasonable, if there is just a small risk for cancer, mammectomy would be declined - decision is less clear between these ends of the scale and at one point, expected benefits and risks of treatment will be equal [121]. This concept of clinical “trade-off” is a key concept in decision analysis and is expressed as “Probability Threshold” which is the most unfavourable ratio of a true positive to a false-positive result that would still be acceptable for the decision-maker [120]. Using this Probability threshold, different strategies can be compared by calculating the “Net benefit” as Net benefit = Benefit − (harm × trade-off), as further described by Vickers et al [121, 122] using the following formula:

$$\text{Net Benefit} = \text{True positive}/N - \text{False positives}/N \times p/1-p$$

with N= total sample size, and p= probability threshold considered acceptable.

Net benefit describes the absolute increase in true positives achieved through the deployed strategy at the respective probability threshold [123, 124] compared to not testing and will range from negative infinity to disease incidence [125]. A Decision Curve then visually compares different strategies across different possible trade-offs by putting benefits and harms on the same scale. It thus helps make informed choices on a reasonable cut-off to balance risk and harms [120].

For prediction models that provide a probability of a positive outcome for an individual case, a probability threshold has to be chosen above which a result will be considered positive. In Decision Analysis, this threshold aligns with the probability threshold of the strategy. A Decision Curve hence can visualize net benefit against different probability thresholds and provide guidance on the best cut-off for prediction models [76, 122, 126, 127].

8.10 Description Of Missing Data And Patterns Of Missingness

Them majority of missing values could be found in the retrospectively assessed data from baseline visits, particularly those requiring laboratory analysis. Only half of the study population (56%) had a complete set of the main baseline assessments such as weight, BMI, CD4 count, Haemoglobin, White Blood Count and Lymphocyte Count, with a large variability across sites, ranging between 95% of clients with complete baseline parameters in Mbeya Referral Hosptial and 19% at Ruvuma Regional Hospital. Liver function tests and Creatinine - that were only required for clients with a suspected liver problem or receiving Tenofovir (TDF) containing regimens respectively- were even more restrictively performed, and not available for any client at some sites. CD4 counts were only missing for 8% of the population and were available for most patients even at sites where other safety variables had not been done. On study visit, the datasets were more complete, with the exception of differential blood count, that had not been performed at some of the sites.

Model.Throughout we performed complete case analysis. However, for some analysis, continuous variables were meaningfully categorized as outlined above with a category including missing data. Only for the Propensity Score Adjustment using the Generalized Boosted Model Technique, where imputation is performed as part of the program, continuous variables were imputed as part of the adjustment process.

8.11 Analytical Methods Considering Clustering And Sampling Strategy

8.11.1 Use And Functionality Of STATA “svy” Command

To generate estimates on the population level that account for study design, the STATA “svy” commands [107, 128] were used to analyse the primary endpoint and sub-group outcome measures.

Through these commands, probability weights are applied and the results are adjusted for design effects and strata. Further, Taylor linearization for variance estimation is used which is appropriate for multiple stages of clustered sampling. The t rather than the more common z statistic are applied to test the significance of coefficients which is relevant in clustered samples with a small number of clusters. In this thesis, p-values based on t statistics and referring to the population-based estimates will be denoted with capital “P” rather than “p”, which will denote statistical tests in the unadjusted data. For model diagnostics, the “svy” command provides an adjusted Wald test [128].

Site stratification by Region and Health Care Level	**Total Eligible sites**	**Total population eligible**	**Sample size** PPS sampling by ART strata	**Wight**
Mbeya Referral Hospital	1	6661	100	66.61
Mbeya Regional Hospital	1	16288	100	162.88
Mbeya District Hospital	9	1922	100	172.98
Ruvuma Regional Hospital	2	2065	100	41.3
Ruvuma District Hospital	5	2108	100	105.4
Rukwa Regional Hospital	1	1410	102	13.74
Rukwa District Hospital	5	1505	100	75.25

Figure 6: Sampling Weights By Sites

In all analysis using the svy command, sites were designated as the first level sampling units. Probability weights as derived from the sampling probability at the site and patient-level as described in section 8.3.1were applied and ART categories were included as strata. Only in the Propensity Score Adjusted regression analysis, probability weights were replaced by the Propensity Score.

8.11.2 Correction Of Standard Error To Account For Sites As Clusters

The study design was in itself self-weighted through site selection and probability sampling, reducing the need for weighting of the sample to adjust for non-equal probabilities of selection. As proportionate stratification applied in stage 2 of the

sampling process usually reduces the standard error resulting in a design factor of less than 1[129], the unadjusted standard error has to be considered the more conservative and hence acceptable to use. Thus some secondary analysis did not include weights or design adjustment. However, in these cases, the vce (cluster clustvar) option of Stata was used to obtain a robust variance estimate that adjusts for within-cluster correlation of participants [130-134].

8.11.3 Specific Considerations Underlying Methods For Predictive Score Development

Although clinical scores are often developed in multicentre trial datasets, methods to account for the clustered data and inherent violation of the unrelatedness assumption are not standardized. In a review of the various clinical predictive scoring system in cardiology, Wynants showed that clustered data is often either ignored or circumvented [75]. Ignoring the cluster effect can lead to underestimated standard errors and effect estimates [75] while accounting for clustered data improves model calibration [135], discrimination and performance [135, 136]. In response, various methods have been used to reflect the clustered study design in predictive modelling, ranging from the inclusion of a centre-level covariate and “leave-one-centre-out”-cross-validation to mixed effects, random effect, logistic or stratified regression models [75, 135].

In this analysis, accounting for clusters was especially important as the sites had shown to be independent risk factors of virologic failure that could not be fully explained by the health care level as strata. Including the site as one of the model variables, on the other hand, was not desirable for a generalizable predictive model, as the value for sites not represented in the study sample would be unknown. To not only mirror the sample design and correct the standard errors to account for the reduced independence of data from clients accessing the same clinic [137], but also to better reflect the impact of site-level influence on treatment outcome [138], we developed a multi-level model as described in section 8.9.4. However, while the methodology would allow investigation of between and within-site variance or across level interactions, the sample size on cluster level was too small to provide sufficient power for such an analysis.

9 Results

9.1 Participants

Between January and December 2013, participants were recruited over 4-8 weeks at each site, starting in Mbeya Region (Mbeya Zonal Referral Hospital, Mbeya Regional and Mbeya District Hospital), Ruvuma Region (Regional followed by District Hospital) and closing with Rukwa Region (Region followed by District Hospital).

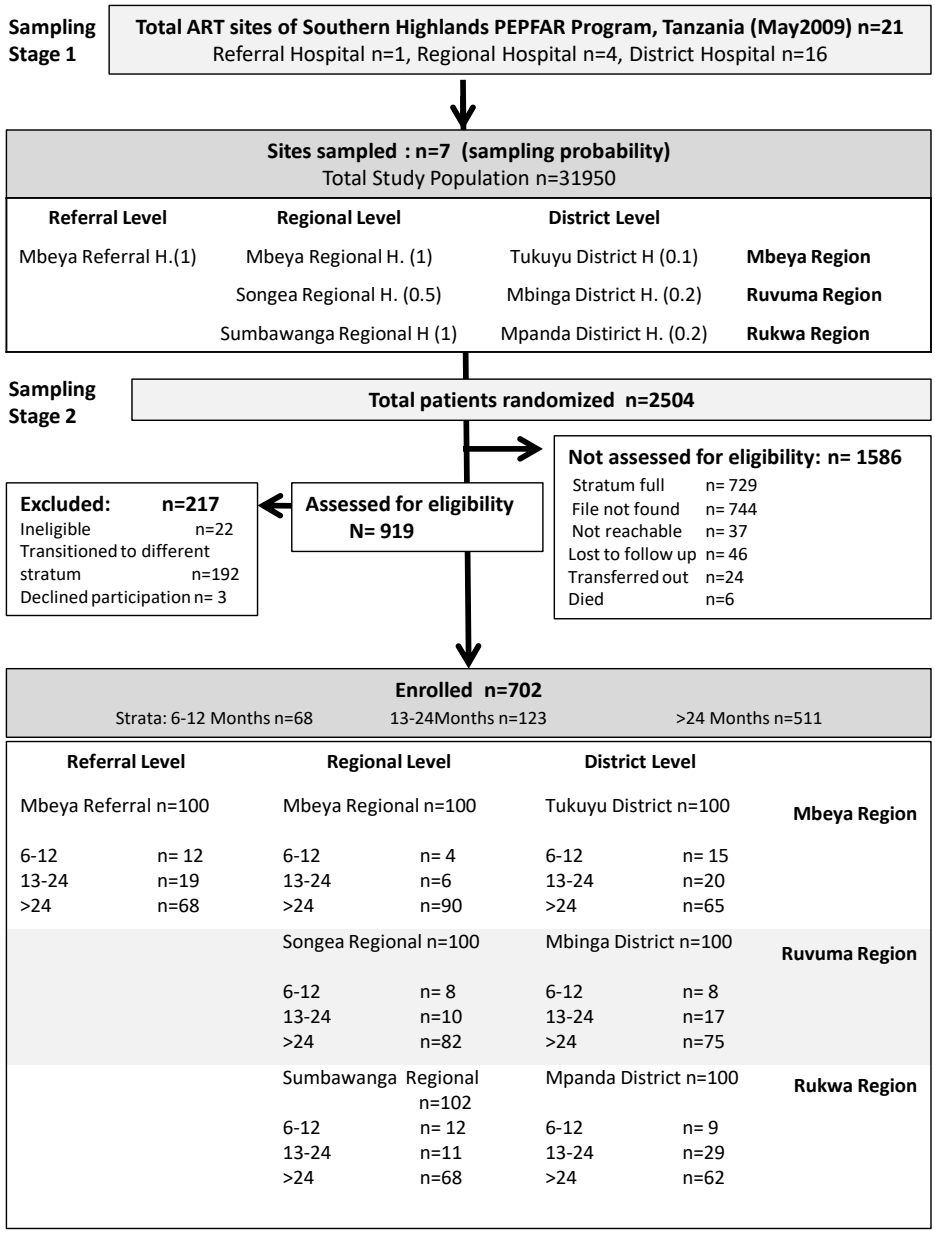


Figure 7: Flow-chart of Study Design And Recruitment Outcome

Of a total of 31950 patients on treatment at all study sites, 2504 patients were pre-randomized. Of these, 1586 were not assessed for eligibility as they had not been reachable, were lost to follow up, died, transferred out or their file had not been found. Of 919 assessed, 192 had changed stratum in the interval between random selection and recruitment, 3 declined participation and 22 were ineligible for other reasons. As two enrolled in the wrong stratum were replaced, An additional recruitment of two participants compensated two of the three subjects with missing viral load information. A total of 702 patients were enrolled and the minimal targeted sample size was met at all sites. All consented patients completed study procedures (Figure 7).

9.2 Descriptive Data

9.2.1 Demographic Data

With 10% in the 6-12 month stratum, 18% in the 12-24 month stratum and 73% in the > 24-month stratum, participants had been on treatment for a median of 48 months at the study visit. The majority of the study population were female married farmers with completed primary education and a median age of 43 years at study visit, who provided for 2-5 persons financially. They lived in a household of a median of 3 adults and 2 children without electricity (Table 1).

9.2.1 Morbidity In The Study Population

Most participants initiated ART at WHO clinical stage 3 (63%), with 9% starting ART at Stage 4 and 6% at stage 1 and 3 participants missing clinical information prior treatment start. Prior ART initiation, weight loss and mucocutaneous manifestations were the main WHO stage 2 disease events reported with a prevalence of 23 and 24% respectively. Nearly half of the population had experienced unexplained prolonged fever (47%) followed by weight loss >10% of body weight (24%) and chronic diarrhoea (20.5%) as WHO stage 3 events. Overall, WHO stage 4 disease prevalence in the population prior ART start was 9%, with disseminated candidiasis (2.1%) and HIV wasting (2.4%) being the main stage 4 defining diseases.

After ART start, 8% of the population developed a WHO stage 2 or 3 event, and 4% a WHO stage 4 disease, which again mainly included weight loss (8%), skin diseases (4%), oral thrush (1%) but also intra- and extrapulmonary tuberculosis (prevalence of 0.1%, 0.4% respectively).

Table 1: Outcome and Demographic Data Of The Study Population By Site

	MZRH		MbRH		MbDH		RvRH		RvDH		RuRH		RuDH		Total	
Total	100	14	100	14	100	14	100	14	100	14	102	15	100	14	702	100
Outcome	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
VS <1000	81	81	92	92	88	88	83	83	88	89	88	88	93	93	613	87
1000 >1000	19	19	8	8	12	12	17	17	11	10	12	12	7	7	86	12
VS400 <400	71	71	91	91	87	87	78	78	84	85	87	87	92	92	590	84
>400	29	29	9	9	13	13	22	22	15	14	13	13	8	8	109	16
VS50 <50	50	50	72	72	65	65	67	67	70	71	64	64	78	78	466	66
>50	50	50	28	28	35	35	33	33	29	28	36	36	22	22	233	33
ART Stratum	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
6-12	12	18	4	6	15	22	8	12	8	12	12	18	9	13	68	100
13-24	19	15	6	5	20	16	10	8	17	14	22	18	29	24	123	100
>24	69	14	90	18	65	13	82	16	75	15	68	13	62	12	511	100
	Med	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**
Age	44	9	41	10	44	11	44	10	41	9	43	10	43	11	43	10
Gender	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Female	61	61	60	60	54	54	74	74	71	71	70	69	66	66	456	65
Male	39	39	40	40	46	46	26	26	29	29	32	31	34	34	246	35
Marital status	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Married	44	44	42	42	48	48	45	45	51	51	59	58	52	52	341	49
Widowed	37	37	36	36	31	31	27	27	23	23	30	29	24	24	208	30
Other	19	19	22	22	21	21	28	28	26	26	13	13	24	24	153	22
Literate	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
No	8	8	11	11	13	13	11	11	15	15	13	13	33	33	104	15
Yes	92	92	89	89	87	87	89	89	85	85	89	87	67	67	598	85
Education	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
None	20	20	20	20	34	34	23	23	15	15	36	35	41	41	189	27
Primary compl.	62	62	66	66	59	59	66	66	72	72	54	53	49	49	428	61
> primary	18	18	14	14	7	7	11	11	13	13	12	12	10	10	85	12
Profession	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Worker	10	10	6	6	1	1	3	3	6	6	42	41	22	22	90	13
Farmer	32	32	56	56	86	86	80	80	72	72	38	37	59	59	423	60
Business	58	58	38	38	13	13	17	17	22	22	22	22	19	19	189	27
Smokes/drinks	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Yes	18	18	24	24	19	19	13	13	17	17	47	46	16	16	154	22
	Med	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**
# Adults in HH	2	2	1	1	2	2	2	1	2	1	2	9	2	2	2	4
# Children in HH	1	1	2	2	2	2	2	1	2	2	2	2	2	1	2	2
Financial dependents	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
<2	7	7	22	22	8	8	6	6	9	9	26	26	12	12	90	13
2-5	64	64	65	65	76	76	71	71	77	77	69	68	81	81	503	72
>5	29	29	13	13	16	16	23	23	14	14	7	7	7	7	109	16
Electricity in HH	35	35	38	38	9	9	38	38	13	13	36	35	24	24	193	28
HIV status of spouse	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Don’t know	19	19	15	15	17	17	4	4	11	11	9	9	20	20	95	14
I know status	44	44	38	38	43	43	41	41	44	44	50	49	49	49	309	44
N/A	37	37	47	47	40	40	55	55	45	45	43	42	31	31	298	43
Spouse HIV neg.	12	12	11	11	6	6	12	12	11	11	12	12	13	13	77	11
Spouse HIV pos.	32	32	27	27	37	37	29	29	33	33	38	37	36	36	232	33
Spouse on ART	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
No	17	24	11	15	10	13	5	6	9	11	9	11	12	18	73	14
Yes	15	21	17	23	29	37	24	29	25	32	29	35	25	37	164	31
N/A	37	52	47	63	40	51	55	66	45	57	44	54	31	46	299	56
Don't know	2	3	0	0	0	0	0	0	0	0	0	0	0	0	2	0

At study visit, 10% of the study population presented with a stage 2, 6% with a stage 3 and 3% with at least one stage 4 disease (Table 4).

In the median, 2 (IQE 1-3) new onsets of WHO relevant diseases were recorded for each participant, which were mainly found prior treatment initiation. Once ART was started, 75% (n=530) of participants did not develop further diseases, and the morbidity under ART was concentrated on a minority of clients.

Ninety-two per cent (n=643) of the study participants had received a baseline CD4 count, and in the median had 149 CD4 cells/µL. Of those, 82% (n=530) were below 250 CD4 cells/µL as a relevant cut-off for treatment initiation according to the national guidelines at that time, and only three had above 500 CD4 cells at baseline. (Table 2)

Table 2: Clinical Presentation At Treatment Start

	MZRH		MbRH		MbDH		RvRH		RvDH		RuRH		RuDH		Total	
Clinical Presentation at Treatment Start																
WHO stage	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
stage 1	4	4	8	8	2	2	12	12	1	1	14	14	3	3	44	6
stage 2	19	19	18	18	16	16	26	26	19	19	35	34	18	18	151	22
stage 3	62	62	63	63	71	71	58	58	60	60	52	51	77	77	443	63
stage 4	15	15	11	11	11	11	4	4	20	20	1	1	2	2	64	9
Treatment eligibility	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Clin.&CD4	64	64	66	66	75	75	50	50	43	43	49	48	70	70	417	59
CD4 count	32	32	26	26	16	16	35	35	21	21	50	49	22	22	202	29
Clinical Only	4	4	7	7	9	9	15	15	36	36	1	1	8	8	80	11
Unknown	0	0	1	1	0	0	0	0	0	0	2	2	0	0	3	0
Funct. Status	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Working	92	92	93	93	95	95	94	94	70	70	72	71	92	92	608	87
< Working	7	7	6	6	5	5	6	6	30	30	11	11	3	3	68	10
Unknown	1	1	1	1	0	0	0	0	0	0	19	19	5	5	26	4
Vitals	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**
Weight (kg)	57	9	54	11	54	10	55	10	52	10	52	8	53	8	54	10
BMI	22	3	21	4	21	4	22	4	20	4	21	3	20	3	21	4
CD4 count	134	128	131	79	114	91	173	98	179	91	156	85	140	101	149	99
Haemoglob	12	2	12	2	11	2	11	2	11	2	12	2	10	3	11	2
Lymphocyt	2	1	2	3	1	1	2	2	1	1	2	1	2	1	2	2
ALT	20	19	23	27	27	23	22	14	28	12	22	20	16	11	23	21
Creatinine	64	49	70	64			79	36	88	176	185		79	75	74	82

Clinical Failure

Three per cent of the study population had been failing clinically (n=21). Of those 8 had met the criteria prior to the study visit and 19 were identified as clinically failing at the time of assessment. Most WHO stage 4 diseases leading to the diagnosis of

clinical failure were weight loss >10% of body weight (n=12), 4 cases were recurrent Kaposi Sarcoma cases and 3 clients presented with Tuberculosis.

Of the WHO defining disease events that occurred after treatment start and did not meet the criteria for immune reconstitutions syndrome (n=204), the majority (62%,

Table 3: Clinical Presentation at Study Visit by Site

	MZRH		MbRH		MbDH		RvRH		RvDH		RuRH		RuDH		Total	
Clinical Presentation at Study Visit																
RR	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Normal	62	62	63	63	67	67	53	53	50	50	69	68	68	68	432	62
Hypertens.	38	38	37	37	33	33	47	47	50	50	33	32	32	32	270	39
WHO stage	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
stage 1	4	4	8	8	2	2	5	5	1	1	13	13	2	2	35	5
stage 2	19	19	14	14	16	16	25	25	18	18	35	34	19	19	146	21
stage 3	58	58	64	64	69	69	56	56	60	60	50	49	76	76	433	62
stage 4	19	19	14	14	13	13	14	14	21	21	4	4	3	3	88	13
IRS No	98	98	99	99	100	100	98	98	98	98	102	100	99	99	694	99
Yes	2	2	1	1	0	0	2	2	2	2	0	0	1	1	8	1
Clin.failure	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
Yes	3	3	3	3	3	3	8	8	1	1	3	3	0	0	21	3
No	97	97	97	97	97	97	92	92	99	99	99	97	100	100	681	97
	Med	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**
Years ART	4	3	5	2	4	2	5	2	5	2	4	3	3	2	4	2
Years HIV+.	5	3	5	2	4	2	6	2	5	2	5	2	3	2	5	2
Peak CD4	567	319	497	318	227	178	467	250	335	297	436	233	306	254	414	289
CD4 gain from BL	270	231	283	190	241	214	315	230	226	218	223	210	204	194	245	215
BMI	24	4	23	4	22	3	22	4	22	4	23	4	23	3	23	4
CD4 count	402	219	416	216	383	230	502	236	378	210	385	221	349	211	402	223
RBC	4	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1
WBC	4	1	5	1	5	1	5	1	5	1	5	1	5	2	5	1
Haemogl	14	2	14	2	14	2	13	2	13	2	13	1	13	2	14	2
Haemogl%	39	4	41	4	40	5	42	6	43	7	39	5	40	8	40	6
MCV	104	10	104	9	103	8	112	146	107	10	99	10	103	10	105	55
Neutrophils	43	12	48	11	43	12	47	11	53	12	46	12	54	11	47	12
Platelets	233	70	253	77	228	168	268	303	277	73	197	55	314	395	253	214
Monocyt	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Monocyt%	10	3	9	2	9	3	9	5	6	4	8	3	6	2	8	3
Lymphocy	2	1	2	1	2	1	2	2	2	1	2	1	2	1	2	1
Lympho %	42	11	39	10	40	11	42	11	40	11	41	12	40	10	40	11
Eosinophi	0	0	0	0	0	0			7		0	0	1	1	0	0
Eosino%	2	3	2	5	4	5					3	4	16	14	3	5
Basophile	0	0	0	10	0	0					0	0	0	0	0	5
Baso %	0	0	0	0	0	0					0	1	1	1	0	0
ESR	25	24	20	23	27	33	30	26	8	9	25	27	50	35	24	29
Creatinine	56	12	53	22	60	16	45	93	44	60	49	14	55	28	52	45
ALT	19	13	16	14	17	12	14	9	12	10	13	36	14	9	14	17
Weight (kg)	62	10	60	11	59	8	55	11	56	12	57	8	57	9	58	10

While sites reported different proportions of clinical failure (p= 0.04), no significant differences could be observed between regions (p=0.2), health care levels (p= 0.06)

and treatment strata (p= 0.2). Clinical failure was not associated with virological failure at any of the cut-offs used (p= 0.699, p= 0.489, p= 0.885 for VS1000, VS400 and VS50 respectively).

Immunological Reconstitution and Immunological Failure

Prevalence of Immune Reconstitution Syndrome (IRS) was 1% in the patient population, with 8 individuals developing new or recurred WHO stage 4 events during the first 6 months of treatment. Next to two cases of newly developed Esophageal Candiasis and extrapulmonary TB, IRS manifested itself as a case of Toxoplasmosis, Pneumocystis Carinii Pneumonia (PCP), Cytomegalovirus Disease and extrapulmonary Cryptococcosis. None of the subjects who developed symptomatic IRS and remained on treatment developed subsequent WHO stage 4 diseases meeting the definition of clinical failure.

Participants gained a median of 254 CD4 (IQR -187,936) between baseline and study visit. CD4 increase differed significantly by ART stratum (p<0.001) and increased with longer time on ART (>24 months: median 291 [IQR -102, 932], 13-24 months: median 153 [IQR -149,676], 6-12 months: median 139 [IQR -36, 457])

However, 14% met the definition of immunological failure, with 12% meeting one, 2% meeting two and 1 person meeting all three failure criteria. For 13% immunological failure could not be ruled out as the parameters needed for assessment – especially baseline and peak value – were not all available for the specific client. A drop of CD4 counts of more than 50% to peak value was the criteria met by most participants failing (63% n=62), followed by a drop below baseline (28%). 8% of clients did not experience immune reconstitution above 100 CD4 counts. Clinical failure was associated with virological outcome at all cut-offs assessed. Clients failing immunologically were 6 times more likely to have a viral load above 1000 copies/ml (OR 5.8 p<0.001), 4 times more likely to be above 400 copies (OR 3.6, p<0.001) and 2 time more likely to have over 50 copies/ml (OR 1.9, p=0.002).

Nevertheless, sensitivity of immunological failure to predict virologic failure was low with a sensitivity and septicity of 44% (95%CI 33-56%) and 88% (95%CI 85-91%) for VS1000, 34% (95%CI 25-45%) and 87% (95CI 84-90%) for VS 400,and 23% (95%CI 17-29%) and 87% (95%CI 84-90%) for VS50.

Table 4: WHO Defining Disease Events And Their Prevalence In The Study Population

	never had		ever had		before ART start		after ART start		ongoing at SV	
Disease Event	n	% Prev	n	% Prev	n	% Prev	n	% Prev	n	% Prev
Generalized Lymphadenopathy	694	94.4	8	5.6	7	1.0	0	0.0	0	0.0
Minor Mucocutaneous Manifestations	506	72.1	196	27.9	163	23.2	33	4.7	31	4.4
Weight Loss $\leq$ 10% of Body Weight	524	74.6	178	25.4	170	24.2	7	1.0	29	4.1
Recurrent Upper Respiratory Track Infections	558	79.5	144	20.5	133	18.9	11	1.6	15	2.1
Herpes Zoster within last 5 years	617	87.9	85	12.1	78	11.1	7	1.0	0	0.0
Unexplained Prolonged Fever >1 month	371	52.8	331	47.2	330	47.0	1	0.1	2	0.3
Weight loss >10% of Body Weight	524	74.6	178	25.4	167	23.8	10	1.4	28	4.0
Unexplained Chronic Diarrhoea >1 month	550	78.3	152	21.7	144	20.5	8	1.1	6	0.9
Pulmonary Tuberculosis	618	88.0	84	12.0	66	9.4	18	2.6	1	0.1
Severe Bacterial Infections	625	89.0	77	11.0	69	9.8	7	1.0	3	0.4
Oral Candidiasis	643	91.6	59	8.4	48	6.8	11	1.6	7	1.0
Anaemia	659	93.9	43	6.1	39	5.6	4	0.6	4	0.6
Oral Hairy Leucoplakia	698	99.4	4	0.6	4	0.6	0	0.0	0	0.0
HIV Wasting Syndrome	674	96.0	28	4.0	17	2.4	11	1.6	12	1.7
Disseminated Candidiasis	684	97.4	18	2.6	15	2.1	3	0.4	0	0.0
Tuberculosis, extrapulmonary	687	97.9	15	2.1	7	1.0	8	1.1	3	0.4
Kaposi’s Sarcoma	691	98.4	11	1.6	9	1.3	2	0.3	4	0.6
Pneumocystis Carinii Pneumonia	691	98.4	11	1.6	10	1.4	1	0.1	0	0.0
Cryptococcal, extrapulmonary	694	98.9	8	1.1	5	0.7	3	0.4	0	0.0
HIV Encephalopathy	697	99.3	5	0.7	5	0.7	0	0.0	0	0.0
Toxoplasmosis	699	99.6	3	0.4	2	0.3	1	0.1	0	0.0
Cryptosporidiosis with Diarrhoea >1 month	701	99.9	1	0.1	1	0.1	0	0.0	0	0.0
Cytomegalovirus Disease	701	99.9	1	0.1	0	0.0	1	0.1	0	0.0
Mucocutaneous Herpes simplex >1month, or any visceral	701	99.9	1	0.1	1	0.1	0	0.0	0	0.0

9.2.2 Antiretroviral Drug Regimen And Co-medications

When starting ART, 59% of clients met clinical and immunological criteria for treatment start, 11% started due to a low CD4 count alone and in 29%, the clinical presentation was the main criteria. The majority of clients started in the WHO clinical stage 3 (63%), 9% started in stage 4, while 6% were at stage 1 at treatment start.

Almost all participants received a Non-Nucleoside (NNRTI) based regimen with either Efavirenz (EFV, 41%) or Nevirapine (NVP, 58%), in 92% in combination with Lamivudine (3TC) and Zidovudine (AZT). Only 7 per cent of women participating in the study had a lifetime history of PMTCT, of which 67% had received single-dose Nevirapine. The main co-medication was PCP prophylaxis in 37%. Twelve per cent of the total study population were using traditional healing interventions next to their ART either consulting with a traditional doctor, using traditional remedies or both (Table 3).

Table 5: Medication At Study Visit By Site

	MZRH		MbRH		MbDH		RvRH		RvDH		RuRH		RuDH		Total	
NRTI	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
AZT/3TC	99	99	86	86	78	78	99	99	92	92	98	96	97	97	649	93
Other	1	1	14	14	22	22	1	1	8	8	4	4	3	3	53	8
NNRTI	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**	**N**	**%**
EFV	31	31	39	39	34	34	43	43	44	44	46	45	50	50	287	41
NVP	69	69	57	57	63	63	57	57	56	56	56	55	50	50	408	58
Other	0	0	4	4	3	3	0	0	0	0	0	0	0	0	7	1
PMTCT	2	2	3	3	4	4	10	10	4	4	6	6	1	1	30	4
Any prophylaxis	35	35	30	30	81	81	17	17	73	73	19	19	9	9	264	38
PCP Prophylaxis	35	35	30	30	81	81	14	14	73	73	18	18	9	9	260	37
TB Prophylaxis	0	0	0	0	1	1	1	1	0	0	2	2	0	0	4	1
Other Co-Med.	46	46	37	37	81	81	42	42	74	74	26	26	11	11	317	45
Currently on TB	7	7	0	0	1	1	1	1	0	0	2	2	1	1	12	2
	Med	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**	**Med**	**SD**
No. of Pills ARV	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1
Total Pills/ day	2	1	3	1	4	1	5	1	4	1	3	2	2	1	3	2

9.2.3 Adherence And Interaction With The Health System

Self-assessed adherence was high with 96% not having missed drug pick up in the last 6 months.

While 67% had a treatment supporter, only 12% actively attended support groups. Disclosure in the private space was high with 85% of all married participants knowing the HIV status of their spouse, of which the majority was HIV positive as well.

Table 6: Adherence and Patient-Clinic Interaction

	MZRH		MbRH		MbDH		RvRH		RvDH		RuRH		RuDH		Total	
	N	%	N	%	N	%	N	%	N	%	N	%	N	%	N	%
Distance to clinic																
0-5kn	43	43	55	55	25	25	71	71	52	52	54	53	51	51	351	50
5-10km	26	26	24	24	40	40	13	13	26	26	30	29	21	21	180	26
>10	31	31	21	21	35	35	16	16	22	22	18	18	28	28	171	24
Clinic access																
Own transport	12	12	11	11	16	16	24	24	43	43	27	27	22	22	155	22
Hired two-wheel	1	1	1	1	11	11	34	34	50	50	27	27	52	52	176	25
Public transport	87	87	88	88	73	73	42	42	7	7	48	47	26	26	371	53
Missed clinic visits last month																
Yes	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1
Time to clinic (min)	45	54	43	92	60	52	30	33	45	51	30	64	60	78	45	64
Attends support group	14	14	9	9	44	44	2	2	3	3	12	12	3	3	87	12
Has Treatment Supporter	10	10	56	56	27	27	100	100	95	95	87	85	97	97	472	67
Has any reason to miss ART	32	32	5	5	25	25	1	1	17	17	25	25	20	20	125	18
ART could be missed due to …																
Toxicity	5	5	0	0	0	0	0	0	0	0	0	0	0	0	5	1
Shared medication	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
Forget med.	19	19	3	3	8	8	1	1	8	8	21	21	11	11	71	10
Feel better	3	3	2	2	0	0	0	0	0	0	1	1	0	0	6	1
Too Ill	3	3	1	1	4	4	0	0	0	0	1	1	0	0	9	1
Travel probl	3	3	0	0	2	2	0	0	0	0	0	0	2	2	7	1
out of stock	0	0	0	0	0	0	0	0	0	0	0	0	4	4	4	1
Has no pills	2	2	0	0	1	1	0	0	0	0	0	0	2	2	5	1
Alcohol	2	2	2	2	12	12	0	0	8	8	1	1	2	2	27	4
Depression	3	3	0	0	1	1	1	1	1	1	0	0	4	4	10	1
told to be cured	0	0	0	0	0	0	0	0	0	0	3	3	0	0	3	0
Uses trad. healers	7	7	6	6	7	7	0	0	7	7	27	27	3	3	57	8
Uses trad. medicine	13	13	2	2	9	9	0	0	7	7	42	41	3	3	76	11
Patient Satisfaction	The following has to improve…															
waiting time	61	61	22	22	22	22	0	0	10	10	40	39	87	87	242	35
skills HCW	9	9	1	1	0	0	0	0	0	0	0	0	2	2	12	2
attitude HCW	13	13	1	1	1	1	0	0	0	0	3	3	1	1	19	3
building	4	4	1	1	0	0	0	0	28	28	2	2	88	88	123	18
overall	12	12	1	1	1	1	0	0	7	7	0	0	78	78	99	14

When asked what they considered the main reasons to miss drug intake, simply forgetting drug intake was the main reasons for incomplete dosing, cited by 10% of

the population. Clinic related aspects such as waiting time, staff attitude or bad quality of care were not cited as a reason to miss a drug pickup visit.

After their confirmatory HIV test, clients were enrolled in care within a median of 4 days. For 25% of clients, test and enrolment were performed on the same day, and 75% had enrolled within 16 days after their confirmatory HIV test. In 7 cases, enrolment in care preceded a documented HIV test result by 7 up to 229 days. With a median of 9 days, most clients were eligible for ART when enrolling for care, and 76% met eligibility criteria within the first three months after enrolment. Only 10% of the study population was still not eligible after one year in care.

Clients started on treatment within a median of 11 days after meeting eligibility criteria, with 58% initiating within 2 weeks and 88% within 2 months. Overall, time from the first diagnosis to treatment initiation took a median of 50 days for the whole population, and for those where a CD4 count at baseline was available, about half of the population had a result within one month after ART start.

Half of the population lived within 5 km from the clinic and took a median of 45min and cost approximately one dollar (1000 TZS) to access care. Half of the participants used public transport, and one quarter either had their own means of transport, walked or used hired two-wheel transportation such as a motorcycle or bicycle taxis. The commitment of these resources appeared acceptable to clients, although transport distance and time were cited as reasons for missing clinic visits.

9.2.4 Description Of Site Level Characteristics

In each region, one district and one regional hospital had been included and the only zonal referral hospital - Mbeya Referral Hospital - as an additional site in Mbeya. The catchment area of these hospitals varied widely between a population of 7,296,789 served by the Zonal Referral Hospital and 564,604 people covered by Rukwa District Hospitals. Regional hospitals tended to be in urban areas with a population above 50000 people, while district hospitals were in smaller villages. Most sites had been operating for 8 years.

Across all sites, a total of 41372 patients had ever been enrolled on ART, most at Mbeya District Hospital (22%) and least at Ruvuma District Hospital (6.5%). Of the total of 14966 patients receiving ART through the study sites, Mbeya Referral Hospital had the biggest and Ruvuma District Hospital the smallest patient population. Only Mbeya referral and the regional hospitals but no district hospitals

had any patients on second line, with the majority of patients managed at Mbeya Referral Hospital. Overall, the proportion of second line patients was 1.1%, with 3.5% in Mbeya. Absolut attrition was highest in Mbeya District hospital, with 5.4 patients lost per patient retained and lowest in Rukwa Regional Hospital with 1.6 patients lost per patient retained. All sites acted as referral facilities to which lower-level CTCs could refer critical patients and in turn down-referred stable patients to lower-level health facilities. Except for Mbeya Referral Hospital, all sites transferred terminally ill patients to home-based care services, but the scale and intensity of collaboration of these inter-facility referrals differed by site. The integration with other services of the hospital such as TB clinic, outpatient clinic and Antenatal Care for management of HIV positive pregnant women varied across sites with different combinations of integration including stand-alone CTCs in Ruvuma and the Rukwa Regional and Ruvuma District Hospital having achieved substantial integration.

Clinics offered ART service on 3 to 5 days per week and required patients to return for monthly drug pick-up visits. Except for Ruvuma District Hospital, sites did not run support groups. Except for Mbeya District Hospital, all sites reported tracing patients lost to follow up. Meantime between diagnosis and enrolment ($p=0.0001$) and from enrolment to eligibility ($p=0.0001$), eligibility to ART start ($p=0.0001$) and between ART start and first CD4 count ($p=0.0001$) differed significantly between sites.

In the median, CTCs were staffed with 10 HCW (range 7-29) of various cadres, with 1 to 8 medical staff (physicians or clinical officers), 3 to 13 nursing staff (nurses, nurse assistants, student nurses) and between 0 and 4 support staff such as nutritionists, counsellors, peer counsellors, community healthcare workers or others. As all CTCs shared their staff with the hospital and applied different rotation schemes for different cadres, ascertainment of the actual staffing situation was difficult. Physicians often attended to CTCs for a limited number of hours per day, while the nursing staff was more likely to rotate on a monthly or weekly basis. Sites might also have a mixed model with a core staff of often retired nurses or clinical officers continuously staffing the CTC and additional staff that would change over time. At the referral hospital, some of the staff were directly funded through partner organizations to supplement the governmental health workforce. Daily clinical staff to patient ratio across the sites was 1:256 (Range 176-295), being highest at the Referral Hospital (1:158) and lowest in Regional Hospitals (1:296). (Table 7)

With 7 sites only, sample size on the facility level was too low to employ statistical methods, however, what could be seen in the site assessment was that differences did not always follow the same pattern. While some aspects of service provision seemed to differ along health centre levels such as much higher lack of consumables such as laboratory reagents, other aspects followed a more regional distribution: Chemoprophylaxis for Pneumocystis Jiroveci Pneumonia (PCP) for those <200 CD4 cells for example as much as INH prophylaxis recommended in HIV infected individuals was generally less commonly provided. Overall, only 54% of those in need received PCP prophylaxis. However, coverage was especially low with Rukwa, with 27% at Rukwa Regional Hospital and 18% at Rukwa District Hospital respectively, while in the other regions, this shortage only affected district hospitals. Only 2% received INH prophylaxis on study visit, all of which were treated at Mbeya Referral Hospital.

Table 7: Facility Characteristics

Facility Characteristic	MZRH	MbRH	MbDH	RvRH	RvDH	RuRH	RuDH
Health Care Level	Referral	Regional	District	Regional	District	Regional	District
Catchment area	7,296,789	2,707,410	339,157	1,376,891	353,683	1,004,539	564,604
Year ART clinic started	2004	2005	2005	2004	2005	2005	2006
Location	>=50.000	>=50.000	2.500-49.000	>=50.000	2.500-49.000	>=50.000	>=50.000
Patients ever enrolled in ART n (%) N=41372	7868 (19%)	8060 (19.5%)	9462 (22.9%)	6107 (14.8%)	3521 (8.5%)	3682 (9%)	2672 (6.5%)
Patients on ART at facility visit n (%) N=14966	3310 (22%)	2071 (13.8%)	1756 (11.7%)	2636 (17.6%)	1407 (9.4%)	2281 (15.2%)	1505 (10%)
Total on 1st line at facility visit n (%) N=14803	3195 (21.6%)	2048 (14%)	1756 (11.9%)	2614 (17.7%)	1407 (9.5%)	2278 (15.4%)	1505 (10.2%)
Total on 2nd line at facility visit n (%) N=163	115 (70.6%)	23 (14.11%)	0.0	22 (13.5%)	0.0	3 (1.8%)	0.0
Nr. Of patients lost per patient retained	2.4	3.9	5.4	2.3	2.5	1.6	1.8
Average number of patients lost per year of operation	506.4	748.6	963.3	385.7	264.3	175.1	166.7
patients lost as % of current patient population.	15.3	36.1	54.9	14.6	18.8	7.7	11.1

9.3 Main Results – Virologic Outcome

With 3 missing viral loads, the primary endpoint was available for 699 (99.57%) participants.

In line with the objectives of this research, we assessed virologic suppression at three different virologic cut-offs based on the rationale in section 6.1.3. and definition in section 8.8.1. with virologic failure being defined as viral load at study visit above 400 copies/ml (VS400), above 1000 copies/ml (VS1000) or above 50 copies/ml (VS50). In line with the STROBE guidelines, we present outcomes as unadjusted outcomes in the study participants ("Study Population"), estimated for the total WRSHCP population adjusted for study design as specified in section 8.9.2. ("WRSHCP Population") and controlled for pre-treatment differences of the participant population through GMB generated Propensity Score weighting as described in section 8.9.2. ("PS Adjusted Population").

9.3.1 Virological Suppression In The WRSHCP

Across the Study Population, 84% (n=590) had achieved suppression below 400 copies/ml, 88% (n=613) had achieved viral load suppression <1000 copies/ml, and 67% (n=466) below 50 copies/ml (Figure 14). The corresponding estimates for virologic suppression in the WRSHCP Population adjusted for study design was: 86% (95%CI 80-91%) virological suppression below 400 copies/ml, 89% (95% CI 85-92%) <1000 copies/ml and 68% (95% CI 61-74%) for <50 copies/ml (Figure 8).

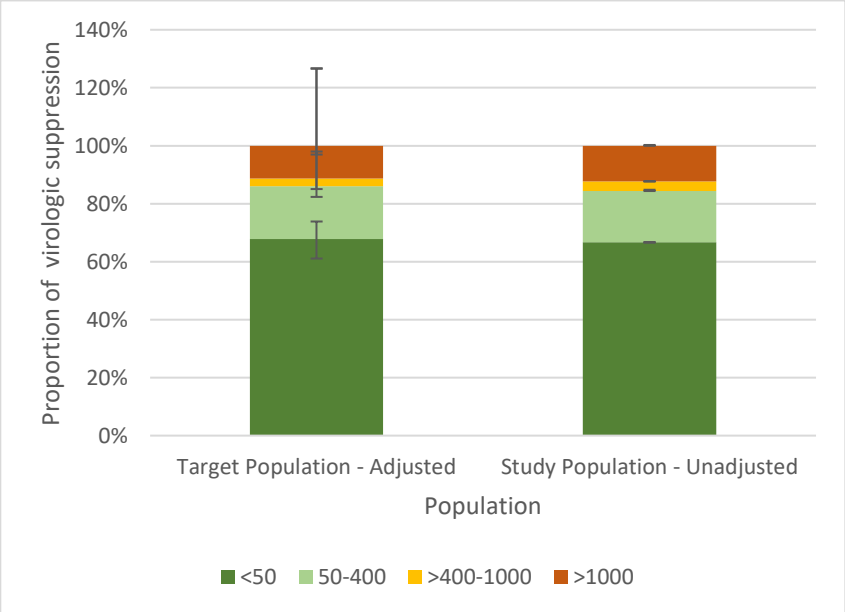


Figure 8: Virologic Outcome by Different Cut-Offs in the Study and WRSHCP Population

9.3.2 Comparison Of ART Strata, Regional And Health Care Level Outcome

Virologic outcome did not differ significantly by treatment strata for all cut-offs used in both the unadjusted study population (Figure 9) and study design adjusted WRSHCP population (Figure 10), although viral load suppression was lowest in the >24 month group with an estimated 87% (95%CI 79-92), 84% (95%CI 74-91%) and 67% (95%CI 58-74%) of the program population below VS1000, VS400 and VS50 respectively.

Likewise, no significant regional differences could be observed and virologic suppression was similar across the regions at a range between 88% (95%CI 82-93%), 86% (95%CI 75-93%) and 65% (95%CI 55-74%) in Mbeya and 92% (95%CI 89-94%), 91% (95%CI 88-93%) and 76% (95%CI 68-82%) in Rukwa for the VS1000, VS400 and VS50 cut-off respectively.

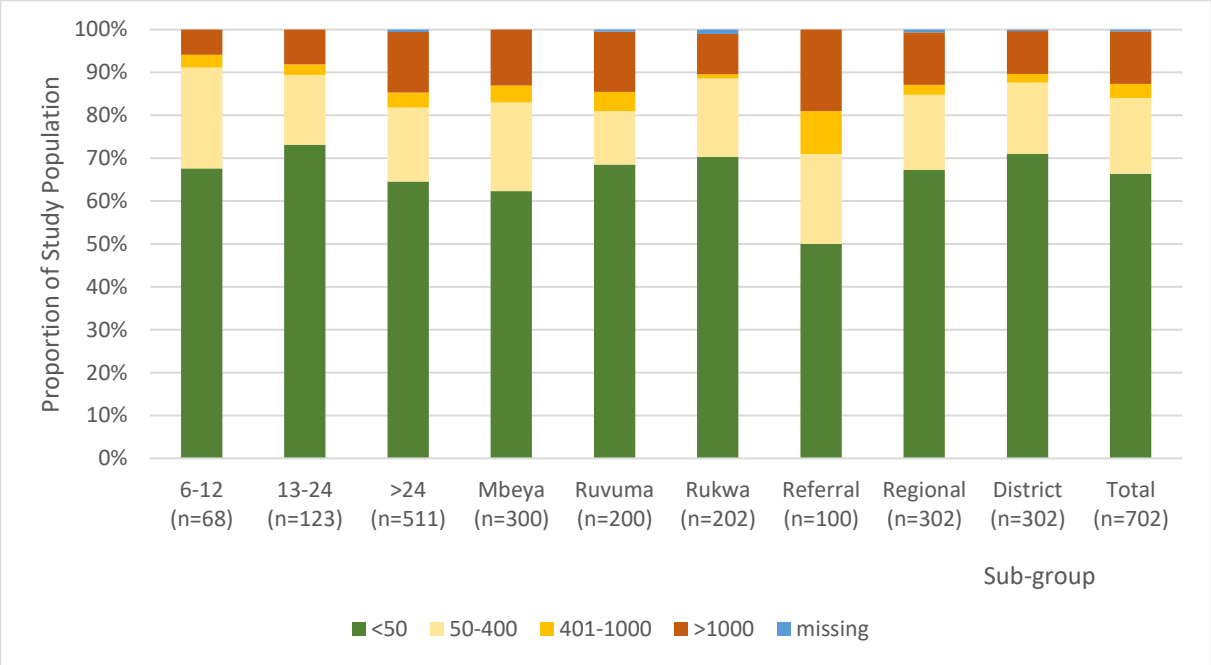


Figure 9: Unadjusted Proportion Of Virologic Failure in the Study Population By Cut-off, Region, ART Stratum, And Health Care Levels

However, when comparing different health care levels, significant differences could be observed for the lower cut-offs (VS50 and VS400) for both study and WRSHCP population (VS50 p=0.003, p=001, VS400 p=0.001, p=0.02 respectively), but not for the highest cut-off of 1000 copies/ml, where groups did not differ significantly (p=0.07, p=0.07 for unadjusted and study design adjusted analysis respectively).

Virological suppression estimated for the program population accessing care at different health care levels was lowest at referral level with 81% (95%CI 81-81), 71 % (95%CI 71-71%) and 50% (95%CI 50-50%) and highest at district level with 89% (95%CI 86-92%), 87% (95%CI 84-90%) and 70%(95% CI 77-61%) for the VS 1000, VS400 and VS50 outcome.

Especially the referral level showed substantially more treatment failure compared to the other levels not only in those meeting the WHO criteria of failing above 1000 copies but also in those showing lower level replication. At the referral level, half of the population was above the test detection limit of 50copies/ml. Without taking population differences into account, the referral level seems to perform much worse than the other health care levels, while the district and regional level seemed to be very similar in outcome across the cut-offs, with virologic suppression at the regional level of 90% (95%CI 83-94%) at VS1000, 88% (95%CI 80-94%) at VS400 and 71% (95%CI 65-76%) at VS50.

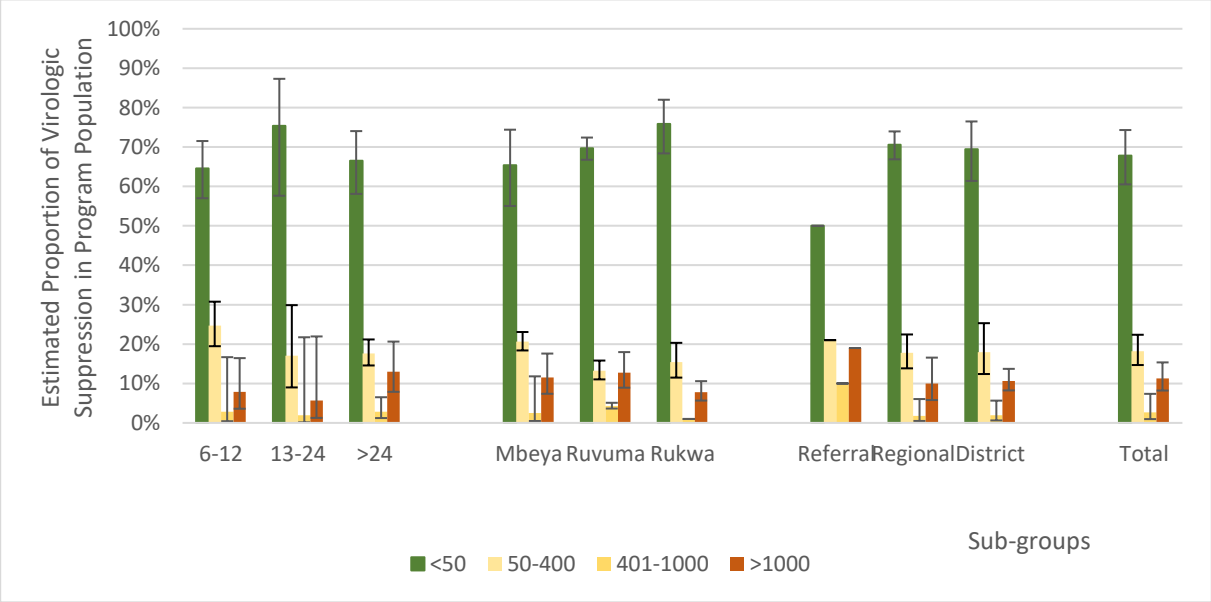


Figure 10: WRSHCP Population Point Estimates and Confidence Intervals For Treatment Outcome By Cut-Off In Sub-groups

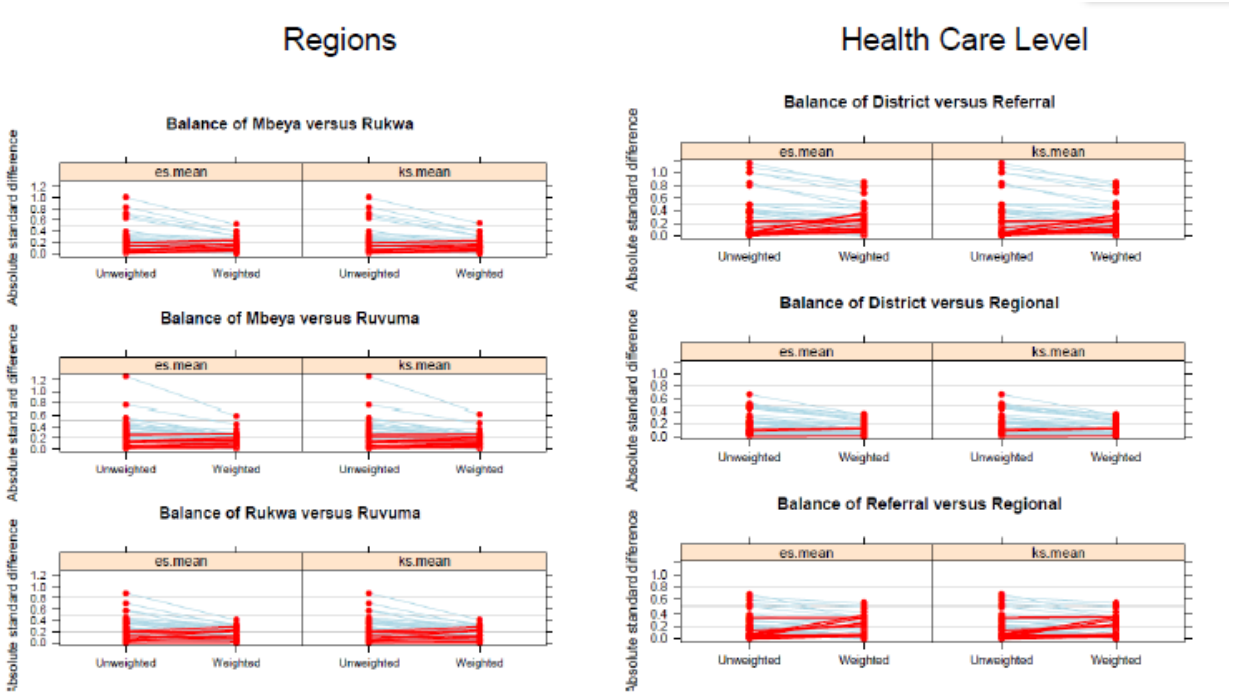
Effect Of Health Care Level And Region On Virologic Outcome Controlling For Pre-Treatment Characteristics In The Study Population.

As population characteristics among regions and health care levels differed substantially, we applied Propensity Score weighted logistic regression described in

section 8.9.2 to assess the impact of health care level and region on treatment outcome controlling for differences in the patient population.

As outlined in the methods section, the “overlap” assumption is crucial pre-requisite for method validity and signifies that there are no values of variables that occur only at one of the sites. For the final PS weights generated for health care level and region, the “overlap” assumption held as demonstrated with the overlap plots provided as additional material in section 13.4. (Figure 35, Figure 33).

Weighting by the PS balanced both regions and health care level, reducing the overall maximum Absolute Standardized Mean Difference (SMD) from 1.3 to 0.6 and 1.2 to 0.9 respectively. As the pairwise comparison of SDM changes presented in Figure 11 depicts, the referral level population was quite different from the lower levels, resulting not only in SMD reduction but also increase after weighting.



Each circle represents the maximum SMD for one of the confounder variables in the weighted or unweighted dataset. Full red circles indicate statistically significant differences, empty circles indicate statistically non-significant differences. Solid red lines indicate differences which are increased through weighting. Results are presented for both stopping rules, SMD and KS statistics

Figure 11: Pairwise Comparison of SDM change for Individual Variables In The PS Weighted and Unweighted Population

Applying the “double robust method” as discussed in the method section, variables that had not been fully balanced by Propensity Score were additionally included as

variables in the logistic regression model. For health care level analysis this resulted in the inclusion of a mode of transport and profession. For the analysis of the impact of region on the outcome, we only included mode of transport and baseline WHO staging relevant weight loss as direct variables as regions were nearly fully balanced following the PS score alone (Figure 11). The results of the logistic regression analysis are presented in Figure 12 and Figure 13. When accounting for differences in the patient population, no differences could be found between the treatment outcome across regions. At the health care level, however, the district health care level significantly differed from referral and regional outcome on all virologic cut-offs evaluated: Participants with comparable characteristics were between 5 to 7 times less likely to fail at district level than at referral level (VS50: OR=0.5, p=0.02; VS400 OR=0.4 p=0.59; VS1000 OR = 0.3, p= 0.003), while for participants managed at the regional hospitals, the observed odds of treatment failure did not significantly differ from the referral level, indicating that the differences observed in the unadjusted analysis were mainly due to the patient population rather than the care received at the respective health care level.

Health Care Level		**Referral**	**Regional**	**District**
VS50				
Risk of failure	Odds Ratio (OR)	1	0.6	**0.5**
	95%Confidence Interval		0.3-1	0.3-0.9
	p		0.058	**0.021**
VS400				
Risk of failure	Odds Ratio (OR)	1	0.6	**0.4**
	95%Confidence Interval		0.3-1.3	0.1-0.9
	P		0.151	**0.049**
VS1000				
Risk of failure	Odds Ratio (OR)	1	0.7	**0.3**
	95%Confidence Interval		0.3-1.4	0.2-0.6
	P		0.254	**0.003**

Figure 12: Estimated Proportion Of Virologic Suppression And Risk of Virologic Failure For Health Care Level Controlled For Patient Population Differences

Region		Mbeya	Ruvuma	Rukwa
VS50				
Risk of failure	Odds Ratio (OR)	1	1	0.9
	95%Confidence Interval		0.6-1.6	0.3-2.5
	P		0.933	0.827
VS400				
Risk of failure	Odds Ratio (OR)	1	1.3	0.8
	95%Confidence Interval		0.361	0.411
	P		0.7-2.3	0.4-1.6
VS1000				
Risk of failure	Odds Ratio (OR)	1	1	0.8
	95%Confidence Interval		0.5-1.8	0.4-1.5
	p		0.902	0.456

Figure 13: Estimated Proportion Of Virologic Suppression And Risk of Virologic Failure For Regions Controlled For Patient Population Differences

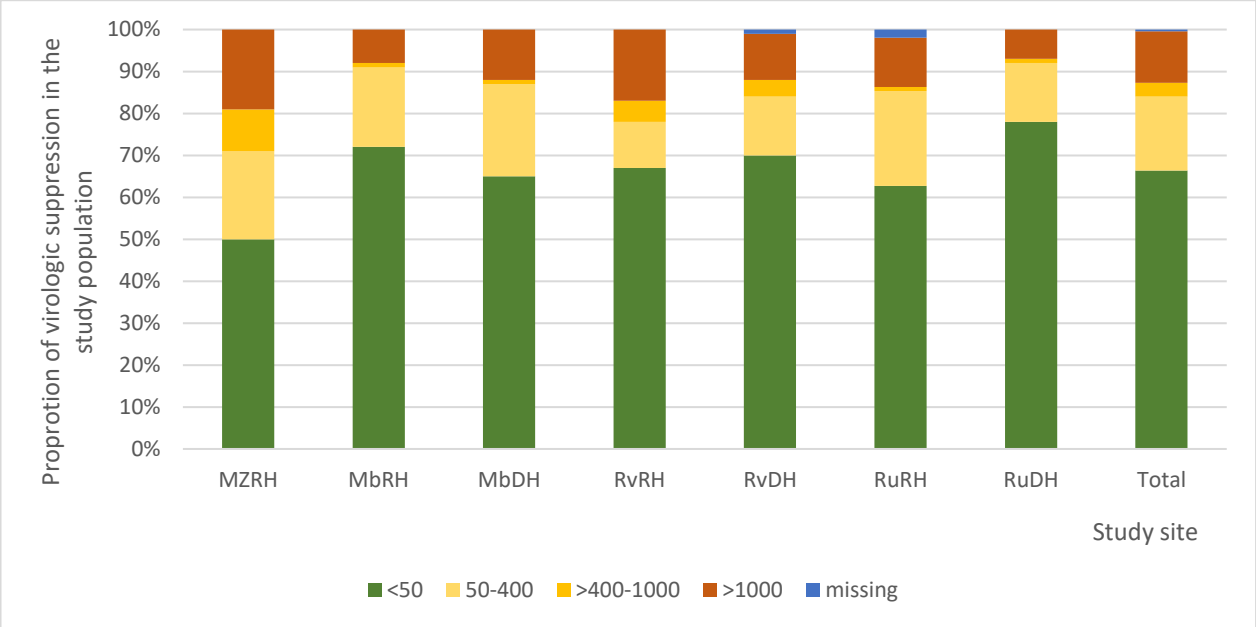
At the district health care level, the “last 90” could be achieved with an estimated 92% (95%CI 90-94) were below the WHO defined cut-off of 1000 copies/ml. All other levels showed lower outcome even when differences in the patient population were taken into account

9.3.3 Virologic Outcome At Individual Study Sites

The Outcome In The Study And WRSHCP Population

At two sites – Mbeya Regional Hospital and Rukwa District Hospital - more than 90% of the population were suppressed below 1000 copies/ml at a study visit, with all other sites showing above 80% virologic suppression. In the unadjusted and study-design adjusted analysis, no statistically significant difference between sites could be found (p=0.1 and p=1 for study and WRSHCP population respectively).

However, the lower virological cut-offs reviled significant differences for both the VS 400 cut-off (p=0.001, p<0.001 respectively for unadjusted and design adjusted analysis) and VS50 (p=0.003, p<0.001 respectively). Mbeya Referral Hospital and Rukwa District Hospital marked the opposite ends of the range with 71% to 92% for VS400 and 50% to 78% for VS50 (Figure 14).



MZRH : Mbeya Zonal Referral Hospital (n=100), MbRH: Mbeya Regional Hospital(n=100), MbDH Mbeya District Hospital(n=100), RuRH Rukwa Regional Hospital (n=100), RuDH Rukwa District Hospital(n=100), RvRH Ruvuma Regional Hospital (n=102) RvDH Ruvuma District Hospital (n=100). Total : (n=702).

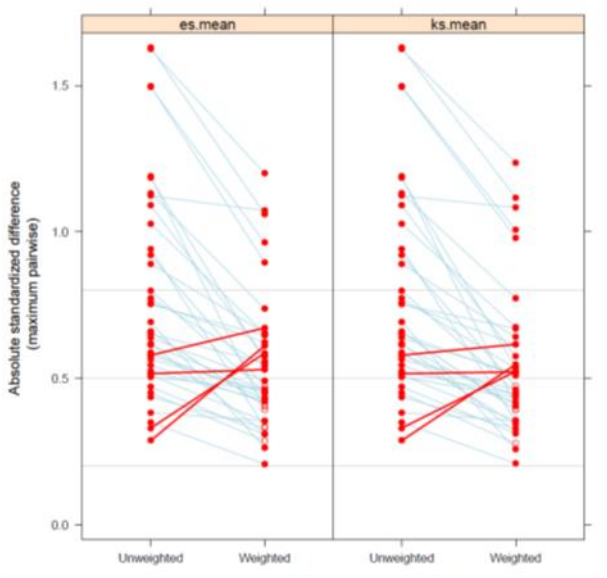
Figure 14: Unadjusted Virologic Outcome in the Study Population By Site and Viral Load Cut-off (n=700)

Effect Of Treatment Site On Virologic Outcome Controlling for Pre-Treatment Characteristics Of The Study Population.

As population characteristics among study sites differed substantially, we analysed the outcome of the different sites adjusting for confounding variables that were considered not influenced by the health care service provided. As described in section 8.9.2, we used logistic regression analysis with a double robust method of adjusting for confounders through Generalized Boosted Model Technique (GBM) generated Propensity Score (PS) weights and direct inclusion of selected confounders that had not been fully balanced by the Propensity Score.

For the final PS weights, the “overlap” assumption held as demonstrated with the overlap plots provided as additional material in section 13.4.,Figure 35. In the balance assessment, PS weights reduced the overall maximum SMDs for the

majority of covariates , with a substantial overall drop of the highest SMD from 1.6 to 1.2. In a pairwise comparison of sites, mean of the SMDs across all variables dropped below 2.5 in all comparisons, but differences remained after weighting the SMD maxima. Those variables (mode of transport to the clinic, profession, baseline WHO stage 3 or 4 changes in body weight and WHO stage 2 Mucocutaneous Manifestations) were included in the final, PS weighted Logistic Regression Model that assessed site impact on the different virologic outcome variables.



Each pair of dots connected by a line represent the maximum SMD for one of the confounder variables in the weighted or unweighted population. Full red circles indicate statistically significant differences of the variable distribution between the sites, empty circles indicate non-significant differences. Blue lines indicate decrease and solid red lines indicate increase of differences through weighting. Results are presented for PS weights generated using the SMD stopping rule (ed.mean) and the KS stopping rule (ks.mean).

Figure 15: Change of Maximum SDM For The Pairwise Comparison Between The Weighted And Unweighted Total Population.

When adjusting for patient population characteristics, differences between the sites became clear at all outcome thresholds:

Mbeya Regional Hospital had the best outcome with a significantly better performance at all three cut-offs. Here, risk of failure directly attributable to the site was between 7 and 8 times lower (VS50 OR 0.3, p=0.003, VS400 OR 0.22, p=0.002, VS 1000 OR 0.3 p=0.03) than at the Mbeya Referral Hospital.

For Mbeya District Hospital, a similar outcome could be observed for the VS1000 and the VS400 cut-off (OR 0.3, p=0.016 and OR=0.24, p=0.001 respectively), but not for VS50. Finally, Rukwa District Hospital had a significantly better outcome in the VS 400 (OR 0.3, p=0.03) and VS50 (OR 0.4 p< 0.03), but not in the VS1000 (OR 0.3 p=0.06) outcome. Both Hospitals in Ruvuma region and the Regional Hospital in Rukwa performed similarly to the Referral Level in all cut-offs used (Figure 16).

		MZRH	MbRH	MbDH	RuRH	RuDH	RvRH	RVDH
Risk of Failure	**OR**	1	**0.3**	0.5	0.6	0.9	0.9	**0.4**
VS50	95%CI		0.2-0.7	0.2-1.1	0.3-1.3	0.3-2.2	0.4-1.9	0.2-0.9
	P		**0.003**	0.071	0.224	0.747	0.738	**0.026**
Risk of Failure	**OR**	1	**0.2**	**0.2**	0.9	0.9	0.8	**0.3**
VS400	95%CI		0.1-0.6	0.1-0.6	0.3-2.2	0.3-3.0	0.3-2.1	0.1-0.9
	P		**0.002**	**0.001**	0.764	0.826	0.601	**0.028**
Risk of Failure	**OR**	1	**0.3**	**0.3**	0.8	0.5	1.0	
VS1000	95%CI		0.1-0.9	0.1-0.8	0.3-2.4	0.1-1.6	0.3-2.9	0.1-1.1
	P		**0.028**	**0.016**	0.706	0.235	0.946	0.066

Figure 16: Estimated Proportion of Virologic Suppression and Risk of Virologic Failure By Study Sites Adjusted For Pre-Treatment Patient Population Characteristics.

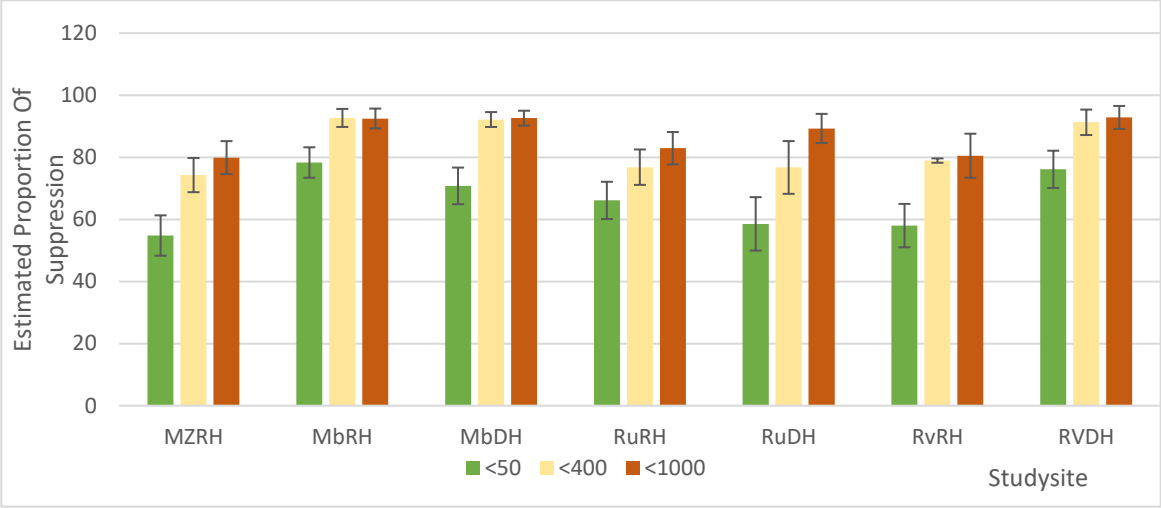


Figure 17: Estimated Proportion of Virologic Suppression By Study Site And Virological Cut-off Adjusted For Pre-Treatment Population Differences.

10 Other Analyses

10.1.1 Patient Factors Associated With Viral Suppression

As described above, three logistic regression models were built to assess patient-level factors associated with virologic failure against the different cut-offs which are presented in Figure 18,

Figure 19 and Figure 20.

The relevance of sites described in section 9.3.3 was confirmed in the assessment of individual factors associated with treatment outcome: Compared to accessing treatment at the Mbeya referral hospital, individuals treated at Mbeya Regional Hospital had a lower risk of virological failure at all cut-offs used (VS50 OR 0.3,

p=0.003, VS400 OR 0.2, p=0.002 and VS1000 OR 0.3, p=0.028) while individual risk of treatment failure at Rukwa District Hospital was lower for VS 1000 and VS 400, but not VS 50 (OR 0.1, p=0.003, OR 0.1, p<0.001 respectively). For Mbeya District Hospital, the Propensity Score adjusted model had predicted a better outcome for all cut-offs compared to Mbeya Referral Hospital, but on an individual level, a significant association between the site and lower failure risk could only be seen for the VS400 cut-off. Similarly, the sites that had not differed in predicted outcome to the Referral Hospital in the Propensity Score adjusted did show association with reduced virologic failure for some, but not all cut-offs. (See Figure 18,

Figure 19, Figure 20 and Figure 16) Next to the study sites, no other variable assessed was retained in all final models for the different cut-offs. Of the demographic variables, age was relevant for the higher cut-offs (OR 0.9, p<0.001, OR 0.97, p=0.009 for VS1000 and VS 400 respectively), but not for VS50, where access of the household to electricity was associated with an increased in virological failure (OR 1.5, p= 0.048).

While clinical failure was not retained in the models for any cut-offs, the occurrence of specific WHO stage defining events - particularly WHO stage 2 or stage 3 weight loss, recurrent upper respiratory tract infections and stage 3 or 4 Candidiasis after ART initiation – increased the risk of individual failure in participants compared to participants who never experienced the event or who had had such a diagnosis prior to ART treatment start. While weight at study visit or baseline was not retained in the models, weight loss of more than 10% of body weight under treatment increased risk of virologic failure for the VS 1000 (OR 5.8, p=005) and VS 400 cut-off (VS400, OR 2.6, p=0.02). Recurrent upper respiratory tract infections meeting WHO stage 2 during ART increased failure risk for the VS 400 cut-off (OR 6.4, p=0.035), while Candidiasis as oral thrush or in disseminated form increased risk of virological failure above 50 copies only (OR 4.4, p=0.04).

The general presence of immunological failure at study visit by itself was retained in the VS1000 and VS50 models as a factor, although it did not meet the significance threshold after CD4 at baseline, peak value and on study visit as much as overall immunological recovery over treatment time were taken into account. Although all final models retained more than one variable evaluating immunological outcome of treatment, only higher CD4 at study visits was significantly associated with less

virologic failure for the VS 1000 and VS400 cut-off (OR 0.994, p<0.001; OR 0.996, p<0.001 respectively) but not with the VS50 cut-off.

> 1000 copies/ml	**OR**	p	95%CI Low	95%CI High
Study Site (Ref.: Mbeya Zonal Referral Hospital)				
Mbeya Regional Hospital	**0.2**	**0.004**	0.1	0.6
Mbeya District Hospital	0.4	0.093	0.1	1.2
Ruvuma Regional Hospital	0.4	0.121	0.1	1.3
Ruvuma District Hospital	0.4	0.170	0.1	1.5
Rukwa Regional Hospital	**0.4**	**0.040**	0.1	1.0
Rukwa District Hospital	**0.1**	**0.003**	0.0	0.5
Age	**0.9**	**0.000**	0.9	1.0
Number Of Alcoholic Drinks Per Week	0.9	0.111	0.8	1.0
WHO Stage 2 Or Stage 3 Weight Loss (Ref: Before ART)				
During ART	**5.8**	**0.005**	1.7	19.9
Never	0.5	0.056	0.3	1.0
Recurrent Upper Respiratory Tract Infections (Ref: Before ART)				
During ART	6.1	0.077	0.8	44.6
Never	0.8	0.547	0.4	1.6
Pulse At Study Visit	1.0	0.145	1.0	1.0
Highest CD4 Count Under Treatment (Ref:<100)				
100-199	0.5	0.344	0.2	1.9
200-300	0.9	0.851	0.3	3.0
>300	1.4	0.561	0.4	4.7
Missing	0.1	0.014	0.0	0.7
Immunologic Recovery Per Year (Ref: Less Than 50 CD4 Cells Per Year On ART)				
>50/y or >350	0.8	0.652	0.3	1.9
Missing	0.2	0.091	0.0	1.3
Immunological Failure (Ref: None)				
Yes	**2.1**	**0.053**	1.0	4.5
CD4 Count At Study Visit	**1.0**	**0.000**	1.0	1.0
Platelets At Study Visit	1.0	0.131	1.0	1.0
Erythrocyte Sedimentation Rate At Study Visit	**1.0**	**0.000**	1.0	1.0
_cons	155.9	0.001	7.0	3472.9
Number of obs 675.0	Log likelihood = -171.8		Prob chi2<0.001	

Figure 18: Association of Individual-Level Factors On Treatment Failure Above 1000 Copies/ml In The Study Population

Here, a higher CD4 count between 200 and 300 at baseline (OR 0.5, p=0.008) and immunologic recovery over 50 CD4 counts per year (OR 0.6, p=0.03) was associated with less virologic failure. Further, an increased Erythrocyte Sedimentation Rate (ESR) as a more general marker of immunologic activity was associated with failure

in the VS 1000 and VS400 but not the VS50 cut-off (OR 1.01, p<0.001, for both cut-offs).

Overall, although a high amount of surrogate parameter for patient adherence and factors acting on adherence were assessed, the only variables in this domain retained in the model described patient satisfaction with building quality and waiting time at the clinic: Dissatisfaction with the clinic buildings was associated with a decreased risk to fail treatment above 50 copies/ml (OR 0.3, p=0.003) compared to those satisfied, while participants who were unhappy with the waiting time were less likely to fail against the VS 400 cut-off (OR 0.4, p=0.01).

>400 Copies(ml	**OR**	p	95%CI Low	95%CI High
Study Site (Ref: Mbeya Zonal Referral Hospital)				
Mbeya Regional Hospital	**0.1**	**0.000**	**0.0**	**0.3**
Mbeya District Hospital	**0.2**	**0.000**	**0.1**	**0.4**
Ruvuma Regional Hospital	**0.3**	**0.003**	**0.1**	**0.6**
Ruvuma District Hospital	**0.3**	**0.004**	**0.1**	**0.7**
Rukwa Regional Hospital	**0.2**	**0.000**	**0.1**	**0.5**
Rukwa District Hospital	**0.1**	**0.000**	**0.1**	**0.4**
ART Stratum (Ref 6-12 Months)				
13-24 Months	1.7	0.343	0.6	5.1
>24 Months	**4.7**	**0.002**	**1.8**	**12.3**
Age	**1.0**	**0.009**	**0.9**	**1.0**
WHO Stage 2 Or Stage 3 Weight Loss (Ref: Before ART)				
During ART	**3.6**	**0.022**	**1.2**	**10.5**
Never	0.7	0.266	0.4	1.3
Recurrent Upper Respiratory Tract Infections (Ref: Before ART)				
During ART	**6.4**	**0.035**	**1.1**	**36.6**
Never	0.8	0.372	0.4	1.4
CD4 Count At Study Visit	**1.0**	**0.000**	**1.0**	**1.0**
Platelets At Study Visit	1.0	0.152	1.0	1.0
Erythrocyte Sedimentation Rate At Study Visit	**1.0**	**0.000**	**1.0**	**1.0**
Waiting Time Unsatisfactory (Ref: Satisfied)	**0.4**	**0.012**	**0.2**	**0.8**
Fear a Reason To Miss Clinic Visit	4.9	0.072	0.9	27.7
_cons	6.3	0.050	1.0	39.9
Number of obs 676	Log likelihood = -228.17598		Prob chi <0.001	

Figure 19: Association of Individual-Level Factors On Treatment Failure Above 400 Copies/ml In The Study Population

>50 copies/ml	OR	p	95%CI Low	95%CI High
Study Site (Ref: Mbeya Zonal Referral Hospital)				
Mbeya Regional Hospital	**0.4**	**0.004**	**0.2**	**0.7**
Mbeya District Hospital	0.6	0.166	0.3	1.2
Ruvuma Regional Hospital	0.7	0.268	0.4	1.3
Ruvuma District Hospital	0.7	0.336	0.4	1.4
Rukwa Regional Hospital	0.6	0.077	0.3	1.1
Rukwa District Hospital	1.1	0.802	0.4	3.0
Number Of Alcoholic Drinks Per Week	1.0	0.146	0.9	1.0
Household Has Electricity (Ref: None)	**1.5**	**0.048**	**1.0**	**2.2**
Candidiasis WHO Stage 3 Or 4 (Ref: Before ART Start)				
During ART	**4.4**	**0.040**	**1.1**	**18.1**
Never	1.5	0.233	0.8	2.9
Has Had Any Skin Infections At Study Visit (Ref: None)	0.2	0.130	0.0	1.6
ART Pill Burden At Study Visit	0.8	0.129	0.6	1.1
CD4 Count At Baseline (Ref: <100)				
100-199	0.8	0.202	0.5	1.2
200-300	**0.5**	**0.008**	**0.3**	**0.8**
>300	1.0	0.964	0.5	1.9
Missing	1.0	0.993	0.3	3.4
Immunologic Recovery Per Year (Ref: Less Than 50 CD4 Cells Per Year On ART)				
>50/Year or >350	**0.6**	**0.034**	**0.4**	**1.0**
Missing	0.9	0.809	0.2	3.0
Immunological Failure (Ref: None)				
Yes	1.5	0.124	0.9	2.6
Unknown	0.9	0.856	0.4	2.1
Platelets At Study Visit	1.0	0.053	1.0	1.0
MCV At Study Visit	1.0	0.121	1.0	1.0
Building Unsatisfactory (Ref Satisfied)	**0.3**	**0.003**	**0.1**	**0.6**
Fear A Reason To Miss Clinic Visit	4.0	0.144	0.6	25.7
_cons	9.6	0.036	1.2	80.2

Number of obs 680 Log likelihood = - -392.54 Prob chi2 <0.001

Figure 20: Association of Individual-Level Factors On Treatment Failure Above 50 Copies/ml In The Study Population

10.2 Development Of A Clinical Score To Predict Virologic Failure

10.2.1 The Theoretical Framework Underlying Model Selection

The theoretical framework developed as a basis for model selection conceptualizes individual treatment outcome as the result of a multifactorial interplay between individual and community factors, with the health care system being considered a

specially relevant area of the community. These dynamics cumulate in two main causal pathways that drive viraemia in the individual:
The Active Drug Pathway influences the concentration of active medication in the blood, while the Immunologic Control Pathway describes the interaction between human host immunologic system and virus.
The Active Drug Pathway is governed by the ability of the individual to maintain constant blood drug levels through regular drug intake which in turn requires regular engagement with the health system for continuous drug supply. At this interface between individual and clinic, factors and clinic dynamics affecting the clinic's ability to provide continuous drug supply and correct prescriptions as much as a welcoming and supportive environment are crucial for individual treatment outcome. The individual's perception of health and illness, any diseases affecting the central nervous system, available resources and a supportive private and community environment are further important factors for this pathway.

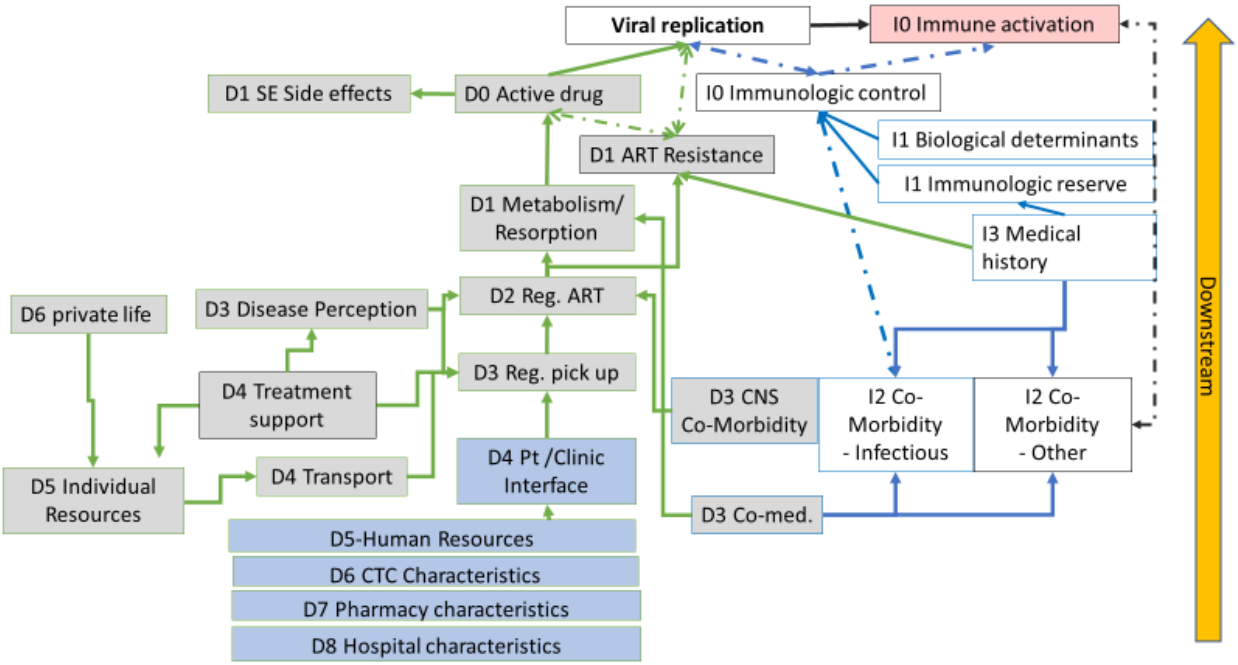


Figure 21:Framework of dynamics leading to virologic outcome

Viral replication in individuals on ART is considered a function of active drug pressure (Dx, green arrows) and immunologic control (Ix, blue arrows) on one side and viral replication and its immune activation on the other. Where factors affect different steps of the pathway, the pathway considered most relevant and most downstream is indicated for clarity. Full lines indicate major unidirectional, causal steps in the respective pathway, while dotted lines indicate important interactions. All steps presented are numbered in relation to their position in the causal pathway starting “downstream” with the direct interaction with the virus. Boxes filled in grey and white contain individual level factors, separated by pathway, blue filling indicates clinic and health care level characteristics.

The Immunologic Control Pathway describes the interaction between the HI-virus and the host immune system. A functional immune system contributes to virologic control, but immunologic pressure can be reduced temporarily in the presence of further co-morbidities. Incomplete viral suppression, on the other hand, leads to chronic inflammatory immune activation, which in turn can cause further co-morbidities and increases cardiovascular risks. While parameters of immune activation are not causal agents of virologic failure, they still can be used as predictive surrogate parameters indicating incomplete virologic suppression (Figure 21). Variables observable at study visit in the majority were more upstream than retrospectively collected variables such as information about medical history and treatment start, as most of the retrospectively collected variable were considered only indirectly impacting viral load mainly through influencing the immunologic recovery capacity.

10.2.2 Characteristics Of The Training And Validation Datasets

Of the total of 702 participants in the RV288d trial, 470 participants were allocated to the training dataset, with 73 failing >400 copies/ml. A total of 232 participants had been allocated to the validation datasets as described in section 8.9.4. The Population Validation Dataset (PVD) included 150 participants with 26 cases of virological failure above 400 copies. The Geographical Validation Dataset (GVD) consisted of all 102 participants managed at Rukwa Regional Hospital including twenty participants shared with the PVD and included 13 cases of virological failure.

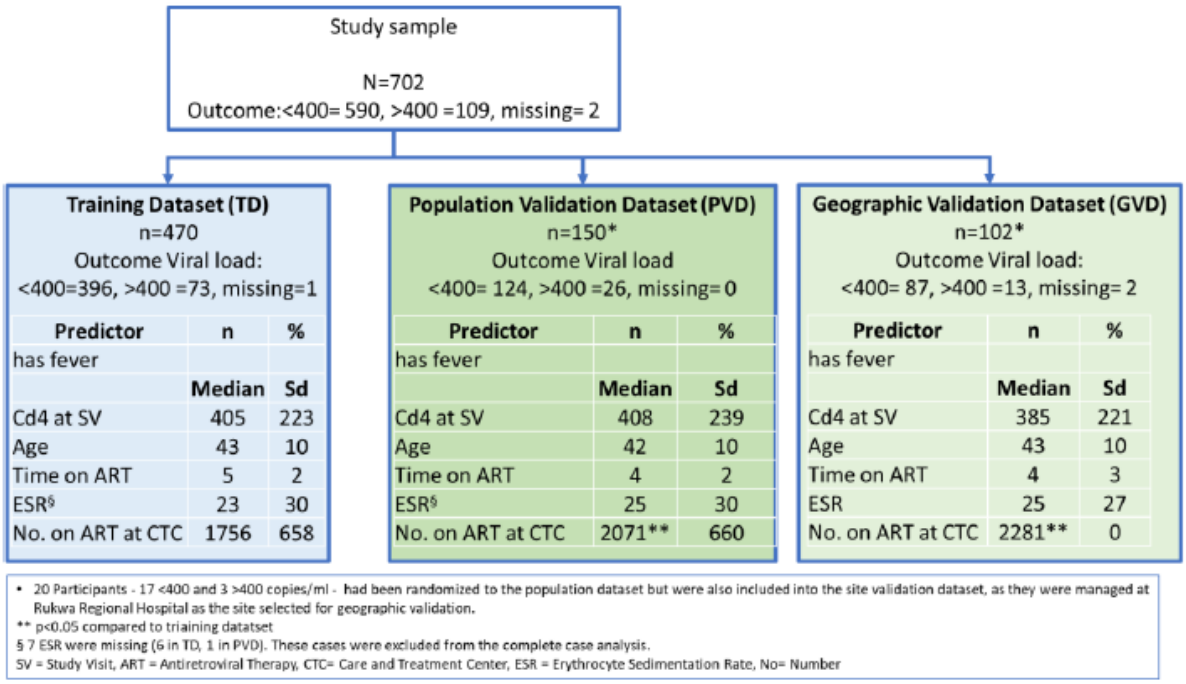


Figure 22: Characteristics Of The Training And Validation Datasets

None of the validation datasets differed significantly in the distribution of outcome variables both at the 400 cut-offs (p=0.6) and around 1000 copies/ml (p=0.3) nor at the Level1 predictors that remained in the final model when compared to the training dataset. However, significant differences existed for the site level (Level 2) predictor (Figure 22).

10.2.3 Model Specification

A total of 200 variables were assessed for model fit, paired with steps in the causal pathway and ranked as described in section 8.9.4. Of these, 85 were excluded due to the high amount of missing values or a low number of events in the total RV288d population. Twenty-eight received rank 3 and were excluded due to considerations surrounding their practicability within the context of the intended use of the model. Sixty-seven variables were considered relevant but less so than another variable

within the same step of the logic model which had received rank 1 paired with this step. Twenty variables (12 level 1 and 8 level 2 variables) were included in the model building process (Figure 23). Details of variable selection are provided in the supplementary material (13.4.2) which includes the rationale for rank assignment, the number of missing values and events in the total RV288d population, the unadjusted association between each candidate predictor and outcome in the training dataset as much as the assigned rank and paired step in the causal pathway.

It should be noted that the drug regimen used would be a desirable variable to be included into the model, as drug regimen are located very much “downstream” in the causal chain and impact of different drugs on treatment outcome has been established in multiple contexts. In our sample, very little variance could be observed between the components of the ART regimen, with most people receiving either EFV or NVP in combination with 3TC/AZT, which is a combination that is being phased out in the public health approach. NVP containing regimen are considered less effective than EFV based regimen, so the NNRTI component was considered relevant for a potential model however, including this variable did not impact model fit hence was not retained. Ideally, a score also would be independent of the regimen used to be robust to regimen changes.

The initially selected Level 1 variables with rank 1 were: time on ART, drinks any alcohol, missed any follow-up visit, mode of transport to the clinic, gender, marital status age, CD4 count, ESR, platelets, respiration rate and WHO T-staging at the study visit. Level 2 variables selected were: clients per clinic day per clinical staff member, ratio clients ever on ART to currently on ART, years of operation of the clinic, number of clients received as referral from lower health care levels, number of clients received as up-referral from lower health care levels, clients on ART at CTC number of clients ever on ART at the clinic and health care level.

	Variable Rank				
Step In The Causal Pathway	1	2	3	4	**Total**
Individual Level variables (Level 1)					
D0 Active drug			1		1
I0 Immunological control	1	3			4
D1 ARV Resistance	1	1			2
D1 Metabolism/Resorption				5	5
D1 Side effects				2	2
I1 Biological determinants	2			4	6

I1 Immunologic reserve	2	12		4	18
D2 Regular ART Intake		1		2	3
I2 Co-Morbidity-Infectious	2	8	10	12	32
I2 Morbidity other		3		4	7
D3 Disease Perception		3		7	10
D3 Regular Pick up visits	1	1		1	3
D3 Co-Medication		2		4	6
D3 CNS	1			4	5
I3 Medical history		11	3	18	32
D4 Treatment support		1	1		2
D4 Transport	1				1
D5 Individual Resources		5	4	1	10
D6 Private Life	1	1	5		7
Facility Level Variables (Level2)					
D4 Pt /Clinic Interface		5	2	9	16
D5 CTC Human Resources	1	4			5
D6 CTC Characteristics	6	3		4	13
D7 Pharmacy characteristics				3	3
D8 Hospital characteristics	1	3	2	1	7
Total	**20**	**67**	**28**	**85**	**200**

Figure 23: Variables Assessed As Surrogate Parameters For Steps In The Causal Pathways Leading To Viraemia And Ranks Assigned

Following the selection process during the model fit described in section 8.9.4., the number of variables was reduced guided by AIC and BIC. Two models were developed with the site as a random intercept. The short model with the most favourable BIC included population mean-centred age, time on ART, squared CD4, log of ESR, and years as fixed level 1 and years of CTC operation as fixed level 2 variable. The long model with the most favourable AIC included additionally the information if the individual had missed any clinic visits and had had any alcohol (Figure 28).

10.2.1 Model Performance

We assessed overall performance through brier score, discrimination with the c-statistics and its graphical equivalent, the ROC-AUC and model calibration visually and through the Hosmer-Lemeshow test.

Apparent performance in the training dataset of both models was similar with a brier score of 0.1 and a ROC AUC of 0.8 (95%CI 0.7-0.9 and 0.8-0.9 in the small and large model respectively) and an estimated model optimism of 0.02 in both models (Figure

28,

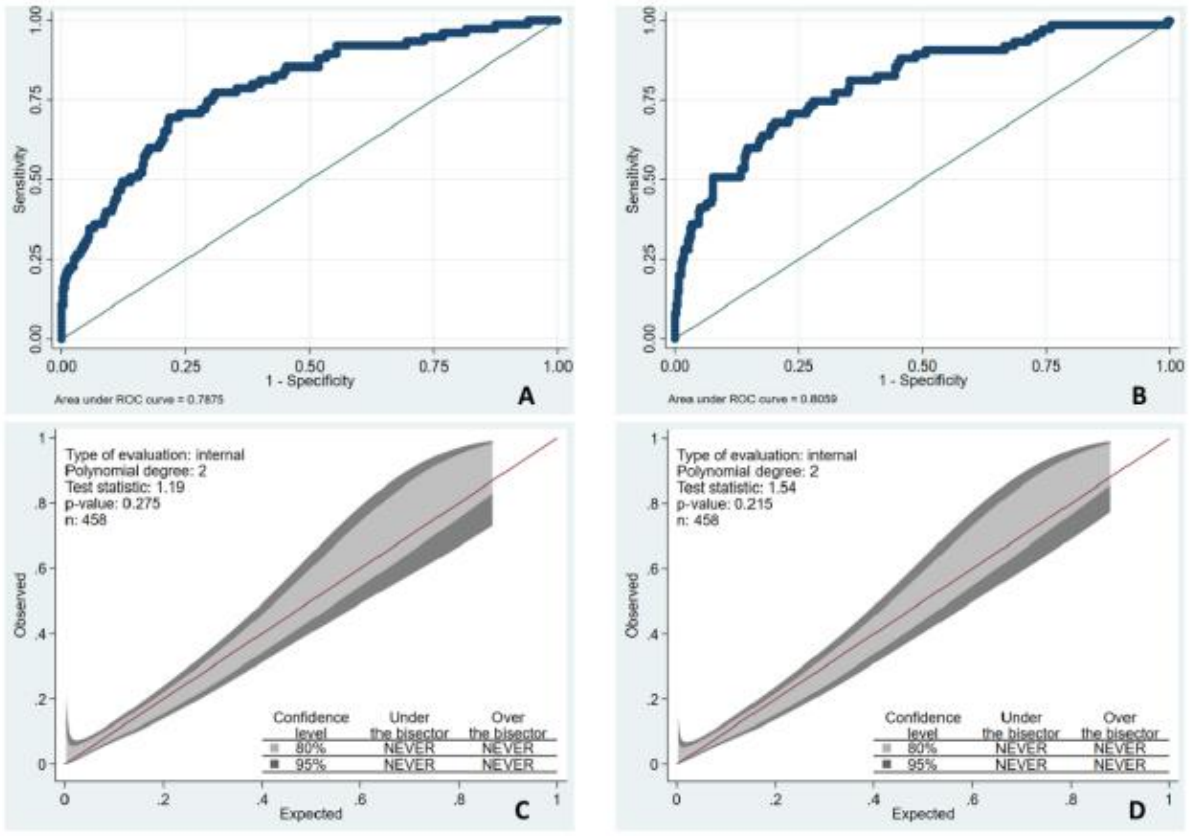


Figure **24**). The models performed slightly less well in the population validation dataset but still comparable with a brier score of 0.12 and a c-index of 0.7 (95%CI 0.5-0.8 and 0.6-0.8 for the large and small model respectively) (Figure 25, Figure 28). Predictive performance in the geographical validation sample was high with a brier score of 0.1 and a c-index of 0.8 (95%CI 0.8-0.9 for both models, see Figure 26, Figure 28).

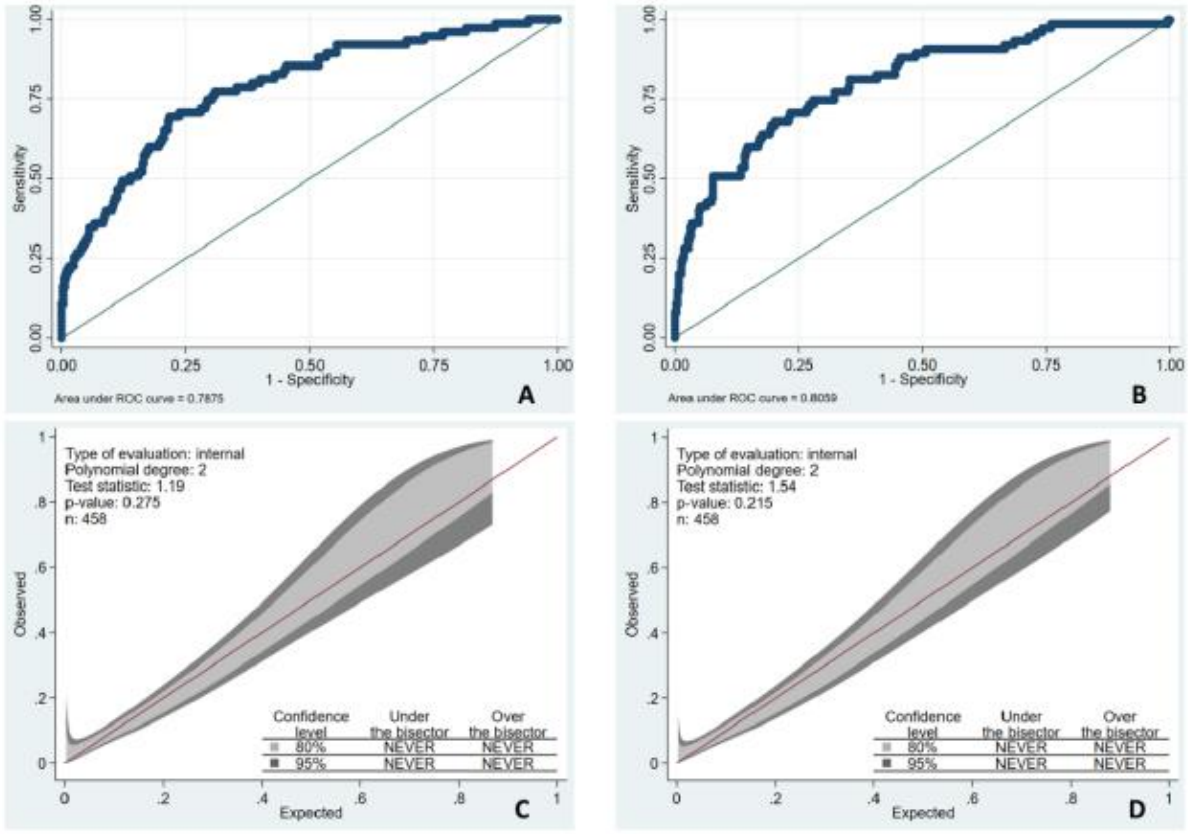


Figure 24: Apparent Performance In The Training Dataset ROC and Calibration Plot of the Small Model (A and C respectively) and the Large Model (B and D)

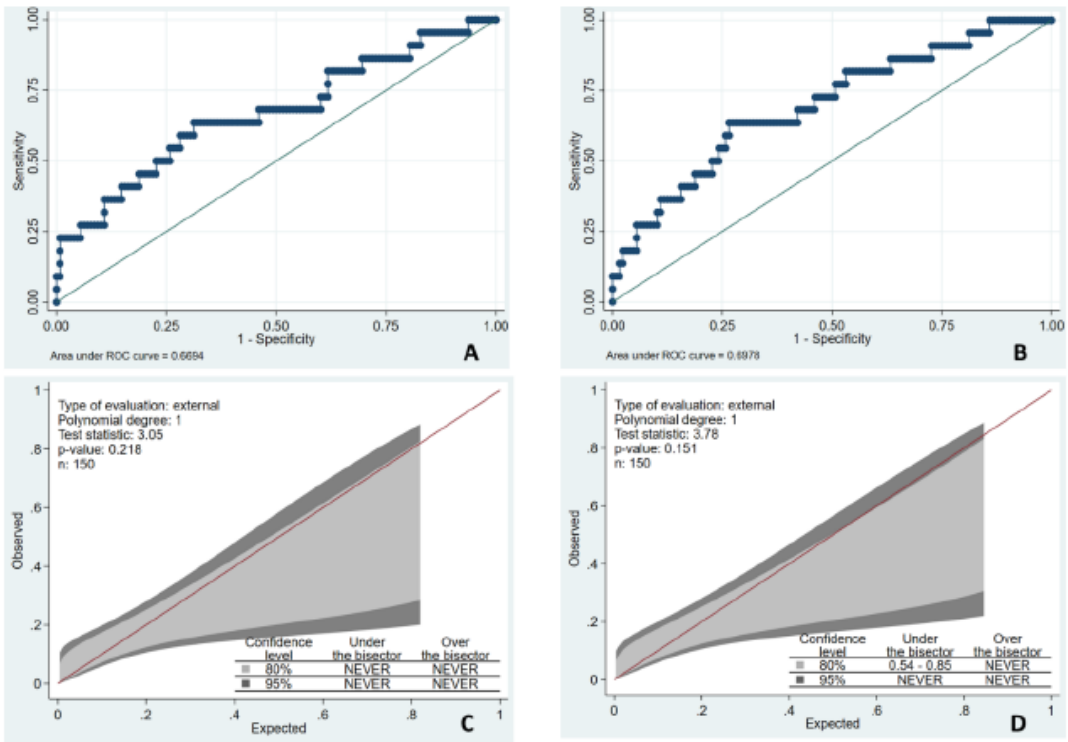


Figure 25: ROC and Calibration Curve for the Small (A and C) and Large Model (B and D) In The Population Validation Dataset

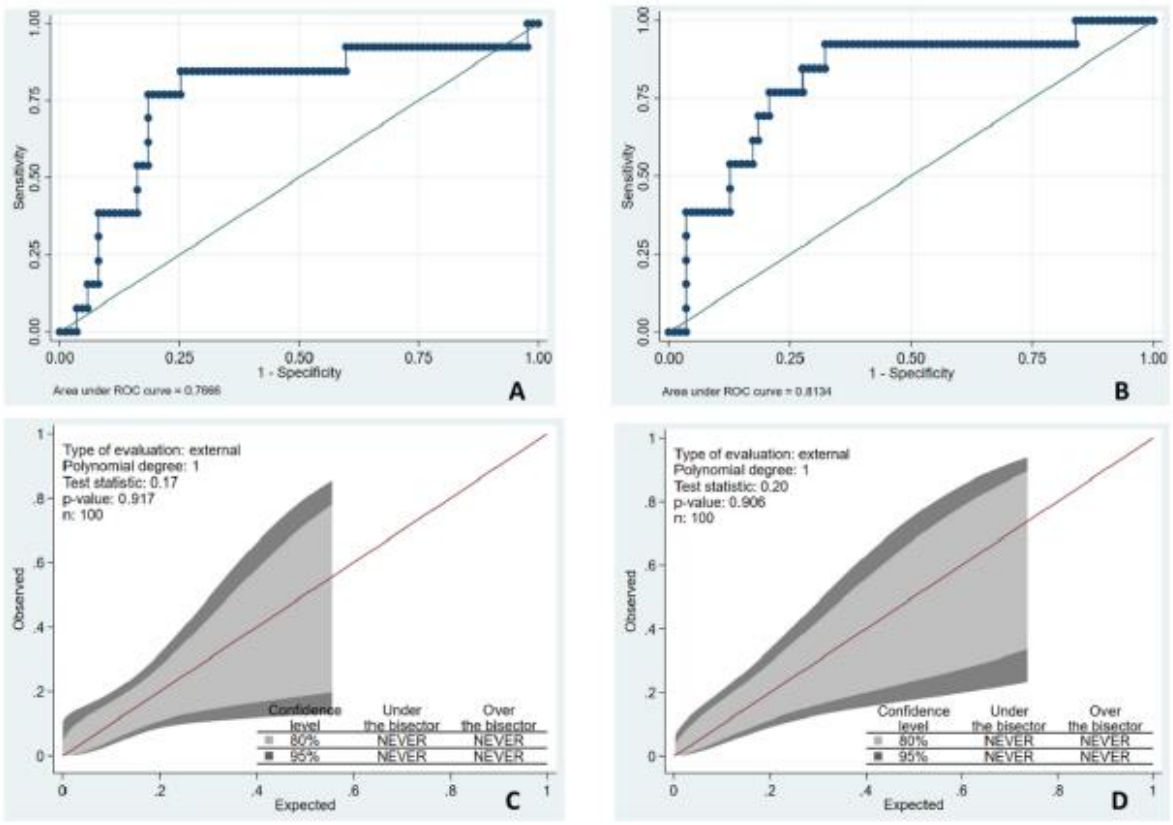


Figure 26: ROC And Calibration Curve In The Geographical Validation Sample for the Small Model (A, C) and Large Model (B, D)

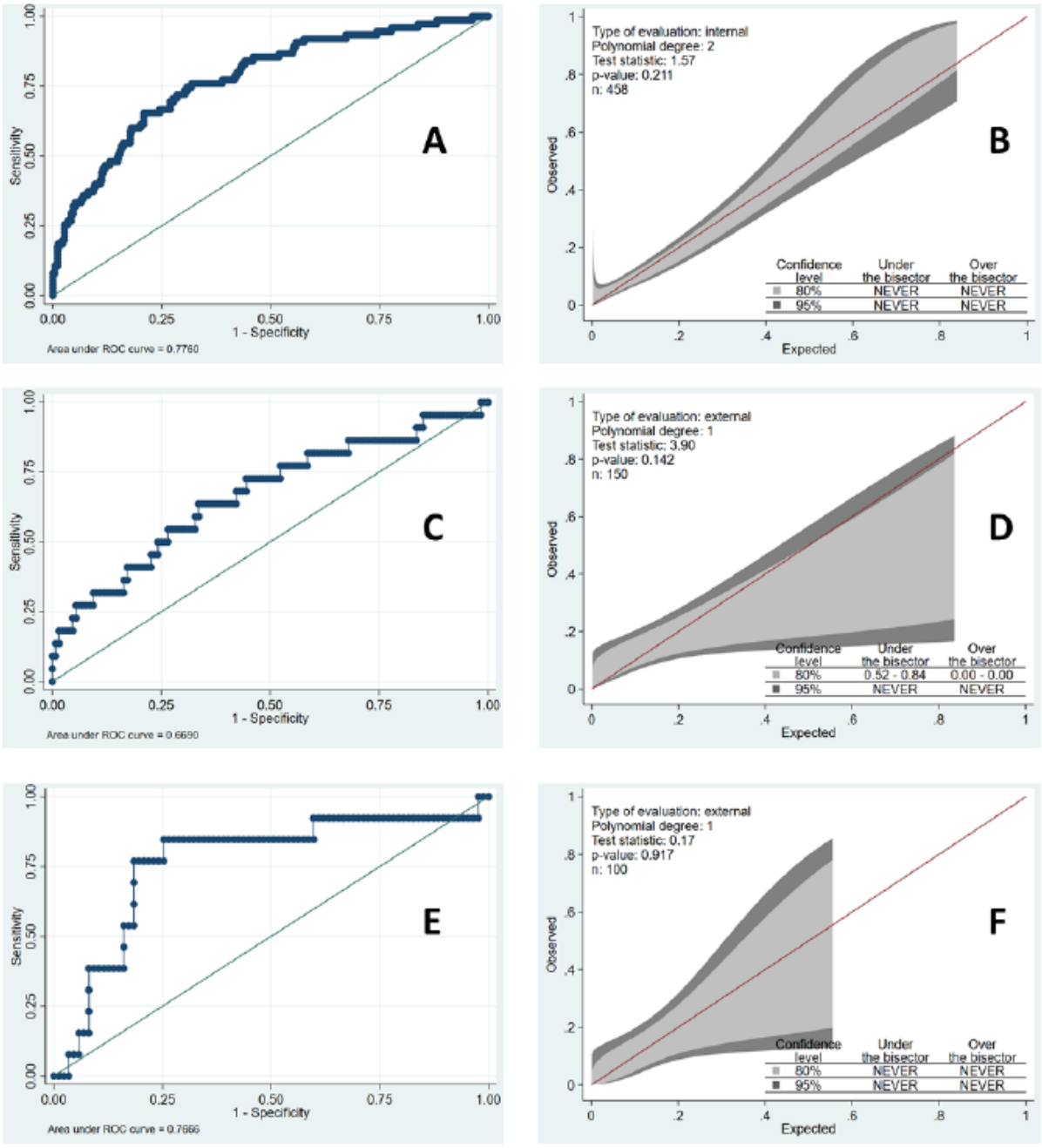


Figure 27: ROC Curve And Calibration Curve For the Logistic Model Used In Constructing The Normogram In The Training Database (A, B), Population Validation Data (C, D) and Geographical Validation Sample (E, F)

	Large Model					Small Model				
				95%CI					95%CI	
Variable	Coef.	SE	p	low	up	Coef.	SE	p	low	Up
Age (centered on population mean)	-0.04	0.02	0.019	-0.07	-0.01	-0.04	0.02	0.009	-0.07	-0.01
Drank any alcohol (Ref. none)	-1.03	0.47	0.028	-1.94	-0.11					
Time on ART (months)	0.22	0.07	0.002	0.09	0.36	0.22	0.07	0.001	0.08	0.35
Missed any follow up visit (Ref. none)	1.35	0.72	0.062	-0.07	2.77					
Square root CD4 at Study Visit	-0.17	0.03	0	-0.23	-0.11	-0.17	0.03	0	-0.23	-0.11
Log ESR at Study Visit	0.45	0.16	0.004	0.14	0.75	0.43	0.15	0.004	0.13	0.73
Years of Operation CTC	1.13	0.27	0	0.59	1.67	1.06	0.26	0	0.54	1.57
_cons	-10.06	2.35	0.000	-14.67	-5.45	-9.56	2.23	0.000	-13.93	-5.18
Site id cons	0.11			0.00	2.72	0.04	0.10		0.00	3.91
Model Performance										
Nr of Obs. (Nr of VL>400 copies/ml)	458 (75)					458 (75)				
ICC site	0.02 (SE 0.03) [95%CI 3E-04-0.45]					0.01 (SE 0.029) [95%CI 0.0001-0.54]				
Optimism in Training Dataset	0.02 (SD 0.06) [95%CI 0.01 - 0.03]					0.02 (SD 0.06) [95%CI 0.01 to 0.03]				
Apparent Performance – ROC-AUC	0.81 (SE 0.03, p<0.001) [95%CI 0.8-0.9]					0.79 (SE 0.029, p<0.001)[95%CI .7-0.9]				
ROC-AUC Population Validation	0.70 (SE 0.06, p=0.002) [95%CI 0.6-0.8]					0.67 (SE 0.07, p=0.006) [95%CI 0.5-0.8]				
ROC-AUC Geographic Validation	0.81 (SE 0.07, p<0.001) [95%CI 0.7-0.9]					0.79 (SE 0.07, p< 0.001)[95%CI 0.7-0.9]				

ICC = Intercorrelation Coefficient, ROC-AUC= Reciever Operating Curve – Aurea Under The Curve, VL= Viral Load, CTC= Care and Treatment Center, ART=Antiretroviral Therapy, ESR = Erythrocyte Sedimentation Rate, Nr of Obs.= Number of Observations.

Figure 28: Coefficients And Model Performance Parameter Of The Two Diagnostic Models To Predict Viral Load Above 400 copies/ml (Virologic Failure) On Study Visit Using A Multi-Level Logistic Model.

10.3 Development Of A Model Nomogram To Use In The Clinical Setting.

To construct a nomogram which could be used in the context of clinical practice in settings without access to electronic prediction tools, the small model was simplified to a logistic regression model accounting for clustered data and re-scaling the continuous variables to represent more relatable units. Performance of this model was similar to the multi-level models but with wider confidence intervals. (Figure 29, Figure 27). The resulting nomogram is presented in Figure 30.

	Small Model for Nomogram				
				95%CI	
Variable	Coef.	SE	P	low	Up
Age	-0.0429	0.0237 4	0.07 1	-0.0894	0.0036 1
Time on ART (month)	0.1997 3	0.0972 1	0.04	0.0092	0.3902 6
CD4 at study visit	-0.0043	0.0016	0.00 7	-0.0075	-0.0012
ESR on Study Visit	0.0127 6	0.0068 4	0.06 2	-0.0006	0.0261 7
Years of Operation CTC	1.1186 9	0.3713 9	0.00 3	0.3907 7	1.8466
Cons	-8.7174	2.0596 4	0	-12.754	-4.6806
Model Performance					
Nr of observation (Nr of VL>400 copies/ml)	458 (75)				
Optimizm in Traning Dataset	0.02 (SE 0.06)[95%Ci .009-0.03]				
Appearant Performance – ROC-AUC	0.8 (SE 0.03, p<0.001)[95%CI 0.7-0.8)				
ROC AUC Population Validation Dataset	0.7 (SE 0.68,p<0.006)[95%CI 0.5-0.8]				
ROC AUC Geographic Validation Dataste	0.8 (SE 0.08, p=0.001)[95%CI 0.6.0.9]				

ROC-AUC= Reciever Operating Curve-Area Under The Curve, VL= Viral Load, CTC= Care and Treatment Center, ART=Antiretroviral Therapy, ESR = Erythrocyte Sedimentation Rate

Figure 29: Coefficients and Performance Measures For Predictive Model Using Logistic Regression Analysis With Sites As Clusters

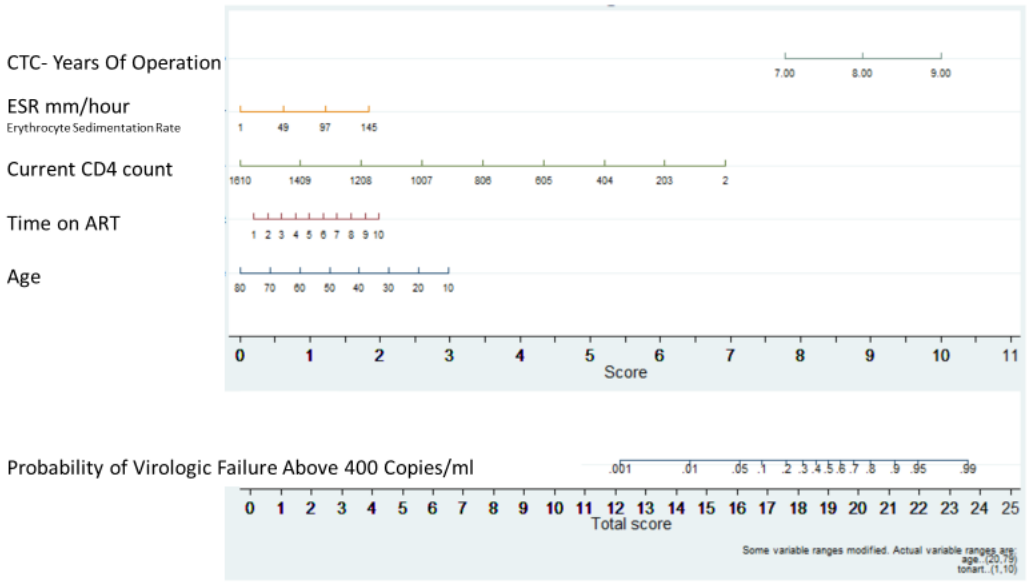


Figure 30: Nomogram To Predict Probability Of Virologic Failure In Patients On Antiretroviral Therapy In The PEPFAR Supported Program In Tanzania.

10.4 Decision Curve Analysis

The method of constructing a Decision Curve has been detailed in section 8.9.4. As summarized in the TRIPOD guidelines: “Decision curve analysis offers insight into clinical consequences by determining the relationship between a chosen predicted probability threshold and the relative value of false-positive and false-negative results to obtain a value of the net benefit of using the model at that threshold.” [92]. In the context of this analysis, the Decision Curve plots the net benefit of different strategies of using viral load tests against different probability thresholds that describe the maximal acceptable number of viral load tests to identify an individual with virologic failure above 400 copies/ml.

The Decision Curve presented in Figure 31: Decision Curves For Different Scenarios In The WRSHCP Figure 31 compares the following different scenarios:

Scenario 1: Test all clients with viral load (routine viral load).

Scenario 2: Test no one for viral load.

Scenario 3: Targeted viral load testing using the clinical and immunological failure criteria only.

Scenario 4: Targeted viral load testing in patients who either fail clinically and immunologically or have a probability of at least >25% to fail as predicted by the nomogram.

Scenario 5: Targeted viral load testing only using the large model (scenario 5a), small model (scenario 5b) and the model supporting the nomogram (scenario 5c).

As direct all scenarios require blood draw – directly for the viral load sample, or for the immunologic and ESR assessments needed for WHO classifications or the model – the direct risk or side effects to the individual in all strategies are comparable. Besides, they are considerably low as they mainly contain the risk associated with a routine blood draw. However, economic considerations around the costs of the HIV viral load test itself as much as the need to set up and maintain the logistics that allow sample transport and result feedback will restrict how many negative viral load tests the program can afford to identify one failing individual.

As can be seen in the Decision Curves presented in Figure 31, if a program cannot afford to test 20 individuals or more to identify one failing individual at a probability threshold of 5%, a model-driven testing strategy (scenario 5) will offer the best net benefit compared to all other strategies. Testing all - (scenario 1) - will only be viable in settings that can afford between 100 and 95% of its beneficiaries, supporting the costs and workload connected to 19 or more negative tests to identify one positive test. If a program cannot afford to test so many individuals, the use of any of the models developed to predict virologic failure (scenario 5) will have preferable odds of finding viraemic patients in those targeted for viral load compared to the alternatives of not testing (scenario 2), targeted viral load testing triggered by clinical and immunological criteria (scenario 3) alone or in combination with a >25% risk of clinical failure in the nomogram (scenario 4).

What is more, scenario 5 is the only strategy that is consistently better than not testing anybody (scenario 2), while the graphs of scenario 1, 3 and 4 all undercut the benefit of scenario 2 (test no one) at some point, resulting in negative net-benefit or harmful impact. For the “test all” approach, this refers to a threshold probability above the prevalence. For the WHO clinical and immunological decision criteria with or without the model (scenario 4 and 3 respectively), this will be at a threshold probability of 32% and 30% respectively. The negative benefit implies that at higher threshold probability, the strategy underestimates the risk of failure.

When comparing the models to each other, their net benefit did not substantially differ across the probability thresholds, although in the higher risk range between

25% and 55%, the model seemed to achieve a slightly higher net benefit than the other two models.

Using the model with a probability threshold in the range between 10% and 30% for targeted viral load testing would identify at least one-third of the true positives. However, higher thresholds would still be beneficiary, especially when using the large model, but less so that with a lower probability cut-off.

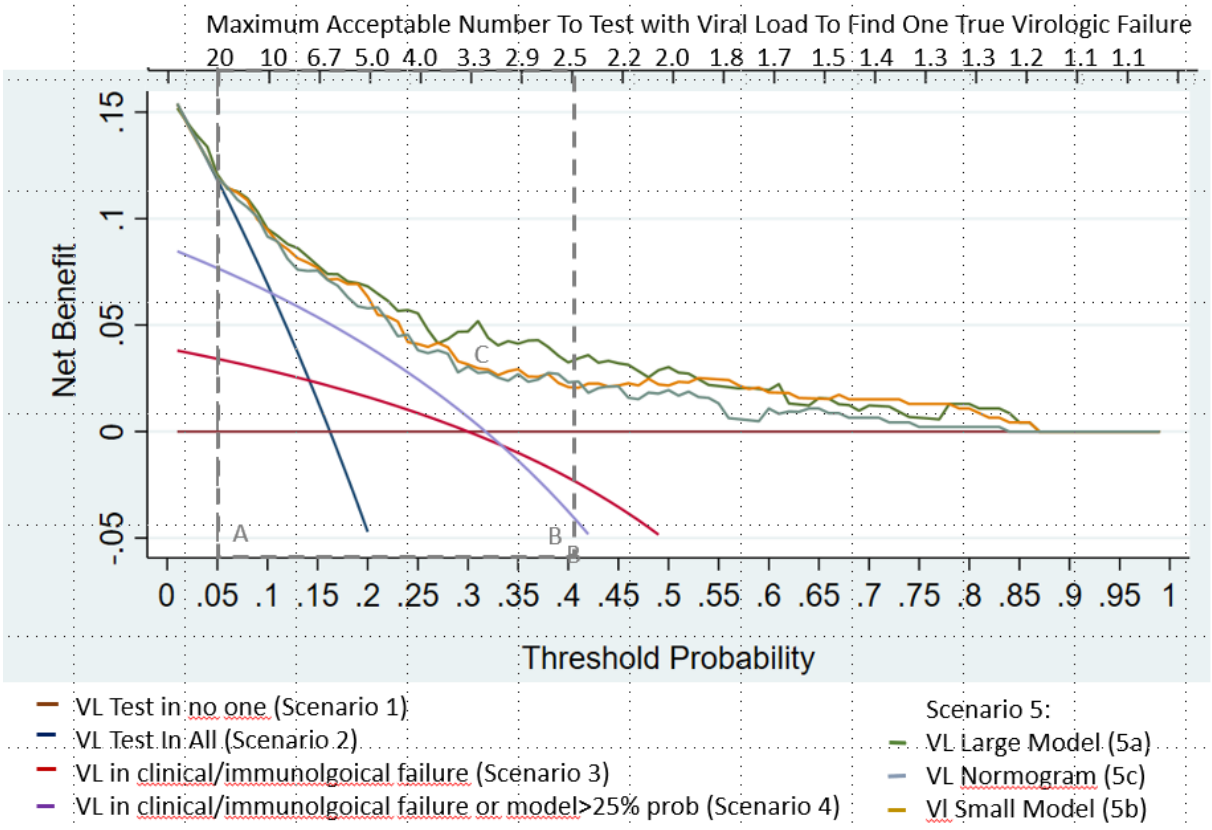


Figure 31: Decision Curves For Different Scenarios In The WRSHCP

Net benefit of different scenarios are plotted against the probability threshold and the corresponding numbers of tests needed to identify one true positive. The area between the dashed lines (A and B) identify the threshold probability range within which targeted viral load testing would be most meaningful, however, even with a higher probability cut-off, model driven viral load test would be preferable over no intervention or clinical and immunological criteria. In the area C, the Large model seems to have a slightly higher net benefit than the other two models.

11 Discussion

11.1 Key Results

This evaluation of the Walter Reed Southern Highlands HIV Care Program (WRSHCP) provided representative information on the population accessing care and treatment in Mbeza, Rukwa and Ruvuma region through the PEPFAR supported national health system in (Objective 1) in section 9.2 and the performance of this system in delivering against the “last 90” as a whole (Objective 2) and disaggregated by region, health care level, ART stratum and site (Objective 3) in section 1.1.:

Using the WHO definition of virologic suppression to determine the “last 90”, we found that 89% (95% CI 85-92%) of the population in care was virologically suppressed as per WHO definition (below 1000 copies/ml) and 86% (95%CI 80-91%) were below 400 copies/ml which was the initial primary endpoint.

No significant differences could be found for the WHO cut-off of 1000 copies/ml across regions, health care levels and sites prior to controlling for pre-treatment population-level differences, but lower cut-offs indicated differences between all health care levels. After adjusting for baseline population differences, the District Healthcare Level consistently had a significantly lower risk of failure compared to the Referral Level which was attributable to treatment at this level, while substantial differences in outcome at the regional and referral level became insignificant.

When comparing sites, 4 hospitals across all regions (both sites in Ruvuma, the referral hospital and Rukwa Regional Hospital) performed similarly, but at Mbeya Regional Hospital, treatment failure was less likely across all cut-offs and the remaining two sites showed differences at the 400 copies/ml cut-off and one additional cut-off assessed. With an OR between 0.2 and 0.3, risk of failure at these sites was 70% lower than at the referral level.

When exploring the impact of participant characteristics on treatment outcome, we compared the treatment strata (Objective 3) and identified characteristics associated with treatment outcome on the individual level (Objective 4) as presented in section 8.9.3. In the comparison of treatment strata, no significant difference in treatment outcome could be detected for all cut-offs, but strata were significantly associated with virologic failure in the regression analysis and time on treatment was retained in the predictive model for treatment failure used to develop the clinical score to predict

risk of failing ART with a viral load above 400 copies/ml (Objective 5). The relevance of sites for treatment outcome was reiterated in this analysis with sites being one of the major factors associated with treatment outcome at any of the cut-offs applied. We developed a clinical score guided by a theoretical framework. This framework conceptualized participant and site characteristics as a surrogate marker for underlying chains of causality that described viraemia as the outcome of host-virus interaction on one side and dynamics that influenced the presence of continuous active drug levels on the other. Details of this framework have been presented in section 10.2.1. Three predictive logistic regression models were constructed, two complex multi-level models with a random intercept and scaled predictors and a single level predictive model using original variables which was converted into a paper-based nomogram. Apparent performance in the training dataset of all models was high with a ROC-AUC of 0.8 [95%CI 0.6-0.9], 0.8 [95%CI .7-0.9] and 0.8 [95%CI 0.7-0.8) respectively. In the validation dataset that included a random sample of participants from the whole RV288d population, the performance was slightly lower with a ROC-AUC of 0.7 [95%CI 0.6-0.8], 0.7 [95%CI 0.5-0.8] and 0.8 [95%CI 0.5-0.8] respectively. External validation in the geographical sample was also satisfactory with 0.8 [95%CI 0.7-0.9], 0.8 [95%CI 0.7-0.9] and 0.8 [95%CI 0.6-0.9].
We explored the clinical utility of these models in Decision Curve Analysis against the WHO routine viral load monitoring approach and targeted viral load strategies using either WHO clinical and immunological failure criteria only or a combination of WHO failure criteria and risk as calculated through the nomogram. When a probability threshold over 5% was applied, all models had a higher net benefit than all comparators, below 5%, the models resulted in the same net benefit than the “test all” approach. Based on this analysis, a cut-off between 10% and 30% failure probability was recommended.

11.2 Limitations

11.2.1 Limitations Of The Study Design

In respect to study design, the following limitations should be considered that might have compromised representativeness of the study sample for the WRSHCP as a whole:

This survey applied a two-stage sampling strategy stratified at both stages, and the completeness and accuracy of list sampling frames as used is crucial for the equal

probability assumption that later on allows generalizability of the findings to the overall population as will be further discussed below [139]. In this survey, sampling frames were constructed using routine programmatic data which had not been previously quality assured other than routine measures applied by the program. In such a situation, it is likely that delay of data entry or other data entry errors resulted in incomplete data frames and unknown exclusion of potentially eligible sites or participants. Possible errors in the sampling frames further might have compromised the self-weighing characteristics of the sample and introduce bias [140].

Information and survival bias inherent in a cross-sectional, time stratified design with retrospective data collection could further be aggravated on Level 1 (Participants) by the implication of our inability to account for 31% of pre-randomized, potentially eligible participants, making attrition a likely positively bias to program outcome. On sampling level 2 (facility level) it needs to be mentioned that while the sampling frame for participant selection was generated not more than 4 weeks prior enrolment at each site, site selection was performed over one year prior to enrolment. This early time point resulted in the exclusion of any sites that might have met the eligibility criteria at the time of data collection. Sites thus excluded might have had a particular service or patient profile leading to specific outcome patterns that might impact overall program results. Further, while this stage was initially designed as probability sampling, the additional sampling frame introduced through the stratification by health care level and the region as much as compromises made due to practical considerations and logistical constraints resulted in some sites being sampled through a method that strongly resembled a purposeful selection [139].

11.2.2 Limitations Of The Analysis

In this thesis, two relatively new methodologies were applied where limitations should be discussed in detail:

Limitations Of The Logistic Regression Models Evaluating Factors Associated With Individual Level Outcome

When discussing the three models using different outcome criteria in parallel, the methodological limitations should be kept in mind: although a standardized and uniform procedure was used to select the models as outlined in 8.9.3, retention of a variable in one model but not in the other does not imply that the variable dropped is not relevant or associated with the respective outcome, but only that in combination

with the other variables in the model it does not contribute to the best model fit in this specific study population. Comparison and interpretation of associations found across models hence have to be done primarily descriptive and with caution. Furthermore, in the backwards selection process, several variables were retained that had very few events in the study population, and thus showed large confidence intervals. However, as they remained to be significant and impacted the AIC and BIC results, they were kept in the model, acknowledging that this might not be a very strong indication of association.

Limitations For The Propensity Score Method And Generalized Boosted Model Technique

As the nature of the study design does not provide for a control group, causal attribution of factors associated with the outcome can generally not be made. However, Propensity Score Methods derived through Generalized Boosted Model Technique (GBM) - which first balances baseline differences and in a second step assesses correlation - is considered to retrospectively generate a baseline balance similar to the balance that would be achieved through randomization, given that no relevant confounder is left out. However, although we collected a magnitude of confounders considered relevant for the study outcome as detailed in section 8.4.2, unknown confounders cannot be fully excluded in such a complex ecosystem and might have affected both the Propensity Score matching as much as the individual-level analysis of factors associated with treatment failure. In the case of confounding, the difference in outcomes may reflect systematic differences in subject characteristics rather than differences of the sites [84].

Limitation Of The Predictive Model

In respect to the developed score, the following limitations should be kept in mind: Due to the low number of events, we were limited with respect to the variables included in the model, hence it is likely that meaningful variables could further increase the predictive capacity of the model, if the number of events would be larger. Similarly, our estimates showed large confidence intervals, which might compromise the value of the score when applying it to any other populations. Although we validated the score internally through a bootstrap method as much as in a population and geographical dataset, both datasets were drawn from the same

underlying population and the use of the score in a different population should be evaluated in further studies.
In the practical use of the score, ideally the score could be used as a point-of- care-test, where the risk of failure could be determined during the clinic visit and adherence counselling or viral load testing could be initiated based on the score. However, the inclusion of the CD4 counts from the day of the study visit will make such a direct application difficult, if no point-of-care CD4 counter is used. While ESR can be determined on-site with very easy procedures, CD4 counts in the majority are analysed in the laboratory and results will be delayed until after the client has left the site. It is unknown how the use of retrospective CD4 count will impact the performance of the model. The implications of this delay are discussed in more detail in section 11.3.4. The limitations relating to generalizability of the score will be discussed in section 11.4. in detail, but it should be noted that while our selection process provided a sample of participants that can speak for the larger program population, some sub-groups of the populations such as transferred-in clients which might have a higher risk of failure have not been covered.

11.3 Interpretation

11.3.1 Virologic Outcome In The WRSHCP

Using [141] the WHO definition of virologic suppression to determine the “last 90”, our findings of 89% (95% CI 85-92%) virologic suppression is in line with other outcomes reported from treatment programs in Tanzania, such as a suppression between 88% and 81% VS <1000 copies/ml observed in Dar Es Salaam [141], a city not covered by the WRSHCP. In a meta-analysis of 184 cohorts from 35 countries, the pooled virologic suppression was 85.6%, which is the lower limit of the estimated 95% Confidence Interval in our findings. Most recent data from the Tanzanian HIV Impact Survey 2017 (THIS) demonstrated 87.7% VS < 1000 copies /ml in those with self-reported ART treatment, which aligns within our estimate [142]. As THIS was a household survey, the comparative outcome between our findings and THIS may indicate that the amount of positive bias in the study is low. It further can indicate that the findings of this WRSHCP evaluation may be representative of the National HIV Treatment and Care Program which is supported through WRSHCP as discussed further in section 11.3.4. When looking at other cut-offs such as suppression below 400 copies/ul as primary outcome of our study, the 86% (95%CI 80-91%) observed

on program level is in line with program outcomes elsewhere [44] : meta-analyses of 49 studies using <500 copies/ml estimated an overall VS of 84% [143].

Regional outcome differences have been reported from other countries [129] [53, 144-146] and have been attributed to both patient-level and health system factors. Regional differences are important when judging the equity of care of an HIV treatment program and its ability to provide equitable access to treatment and care in line with the SDGs to “leave no one behind”. In our analysis, regional differences were not statistically significant, indicating that once an individual is enrolled in care, quality of care is similar across the regions.

However, it needs to be kept in mind that differences related to the location are likely to exist on a sub-regional level. In our study, the majority of the study population lived less than 10 km from their CTC and only 24% travelled more than 10km to access care, while catchment areas of the sites, particularly the Referral Hospital, are much larger. Distance from clinic has been identified as a major factor impacting treatment outcome in other studies [147, 148], but the population represented in the study sample did not indicate that clinic access was a major obstacle to treatment access. Thus, distance to the clinic might either be a prohibitive factor for accessing the health system, or the majority of citizens in the regions outside of the 10km belt access care at lower health care levels such as primary care facilities or community-based programs.

The THIS study indicates that the main challenge in Tanzania is “the first 90”, which refers to HIV testing uptake. In its regional disaggregation, most individuals who were aware of their positive HIV status were on ART, but 60% of the total HIV positive population in Rukwa and 41% ins Mbeya and Ruvuma were not aware of their positive HIV status [149]. Due to this high proportion of unknown HIV Infections, virologic suppression in all HIV positive individuals was 57% in Ruvuma and Mbeya and 43% Rukwa [149], which is far below the 75% targeted by WHO to reduce HIV incidence at the population level. It is possible that equitable regional difference can be achieved, but on a sub-regional level, coverage for both HIV testing and HIV treatment access might be limited. Further research that utilizes geospatial data is needed to explore geospatial dynamics impacting access to care in order to improve not only testing but also treatment coverage.

In this context, attrition should be re-visited as a crucial problem identified in other ART programs. In Tanzania, a comprehensive assessment of attrition showed a loss of 18% of the patient population after 12 months and 36% after 36 months on ART with 10% and 14% mortality at 12 and 36 months of treatment respectively. In this study, mortality and attrition rates both peaked within the first six months [49]. Although RV288 did not directly trace or otherwise account for attrition in the study design, 34% of the individuals earmarked for enrolment on the Pre-Recruitment List could not be identified, and only a few of those could be established as lost to follow up (2%), transferred out (1%) or dead (0.2%). If the remaining individuals are used as a proxy for attrition, the program would be above the <20% attrition threshold targeted by WHO, within range of estimates produced by others [41, 42], but in contrast to the results of the Tanzanian impact survey (THIS) that - as a household survey - collected self-reported ART enrolment at community level. THIS reported that only 5% of those who self-reported as HIV positive were not on ART, while the majority of HIV infected not on treatment were unaware of their positive HIV status [149], which would indicate that the high level of attrition reported from facility level studies and seen also in our recruitment process is either due to unreported death, a documentation problem or a problem of scale. Even if one accounts for the fact that in THIS, self-report and biomarkers did not always align and 10% reporting not to be aware of their positive HIV status had measurable ARV drug concentration in their blood, while 10.0% of those reporting ARV use had no serologic ARV levels, these variances cannot fully explain attrition rates seen in our study.

11.3.2 The Impact Of The Health Care System On Programmatic Outcome

Latest with the adoption of the SDGs, focus on international public health is shifting from disease-specific public health programming to strengthening the health system as a whole. With this shift, the quality and efficacy of service provision come into focus as important factors influencing treatment outcome.

Prior to discussing differences of health care levels and site-specific performance it should be acknowledged that a decentralized health care system can be a cause for regional differences [150]: Regional health offices provide the vertical structures within the health care system responsible for logistics such as procurement of drugs and laboratory reagents. Thus, all sites within a particular region will be similarly affected if these regional structures are compromised and in consequence regional

outcome differences will be observed. In our cross-sectional study with 7 sites, all sites reported drug shortages and difficulties to access essential medication which could be attributed to such regional logistics and procurement structures. To mediate such shortages, sites resorted to local collaboration as borrowing drugs from other sites or by reducing the drug-refill intervals for beneficiaries. The substantial amount of missing values in the area of the retrospectively collected laboratory results at ART initiation may further testify about the incomplete supply of reagents or maintenance of equipment. However, these irregularities did not lead to noticeable differences in regional outcome.

The impact of the health care level on treatment outcome has been explored by others: In Nigeria, secondary satellite facilities showed significantly more deaths in the first 12 months on ART, and less immune reconstitution and virologic suppression in the first 12 weeks compared to the tertiary hospital [151], while smaller facilities elsewhere were associated with a higher loss to follow up and mortality [129]. In contrast, clients at regional and district hospitals in South Africa were less likely to fail treatment than referral hospitals, even after controlling for baseline differences [144]. The latter aligns with our findings that treatment at the district level is associated with a lower risk of failure, supporting the abundance of existing evidence that ART delivered at lower health care levels is at least comparable to higher levels particularly for those stable patients comprising the majority of our study population [51, 65, 152, 153]. However, the categorization by health care level might be too crude to draw meaningful conclusions that can inform program planning. What is more, health care levels are not necessarily defined by a particular set of functionalities a clinic would have but rather relative to other clinics within the respective region, so the comparison is difficult across countries or even between different regions within country.

ART delivery in Tanzania as in other Sub-Saharan countries is increasingly decentralized and fragmented into a variety of ART delivery models [154, 155] that differ in the way service is delivered, their level of integration with other services, their use of digital support tools such as text message supports [156] or other forms of patient tracing [157, 158] or the distribution of responsibility across different cadres of health care workers [155]. Especially with the start of “test-and-treat”, individuals further might move between these models, initiating ART during testing campaigns or community outreach activities and then continuing treatment at a local CTC. The

outcome of a site might not be so much defined by its health care level, but by the service delivery models a site is using and collaborating with. Only a few of these models have been assessed with respect to their impact on virologic suppression [154].

Our analysis at suggest that rather than health care levels, the individual sites are relevant in respect to virologic treatment outcome: While the regional health care level did not perform differently to the referral level as a whole when adjusting for population differences, the regional hospital in Mbeya performed better across all cut-offs used compared to the referral hospital after pre-treatment differences in the patient population were taken into account. Further, two other sites performed significantly better at the 400 copies cut-off and at least one other cut-off used.

It is important to recognize that in the literature various outcome measures are used in parallel from hard endpoints such as mortality, emergency room visits [159] or hospitalization to surrogate parameters such as treatment uptake, engagement in care [154] or conformity with treatment guidelines [160] and other process measures to capture service delivery quality of a facility. Many authors both in resource-rich and developing countries further deliberately use a panel of outcome measures to reflect the quality of care for different sub-groups of patient populations such as adolescents, pregnant women or severely immunocompromised clients [161, 162]. This complicates direct comparison of study outcomes further, especially as delivery models might produce mixed results depending on the outcome measure [69]. Viral suppression achieved within the population of a site is rarely used to describe service delivery quality especially in developing settings. However in the United States, virologic benchmarks have been established as part of the portfolio of outcome measures to assess and compare treatment in different health facilities funded by The Ryan White HIV/AIDS Program, that provides HIV primary medical care and support for uninsured people living with HIV [40]. These benchmarks can be very useful to assess the service at a site as they can describe the impact the service delivered at the site has as a whole, capturing also synergisms between different activities that the site might deploy.

However, our findings highlight that for viral load to be a measure that would allow comparability, it is important to account for population differences. As we could show, binary comparisons between health care levels without adjusting for population

differences can be misleading: while treatment outcome at referral and regional level differed, this difference was not attributable to the health care level but to the differences in the populations treated at these levels. When we adjusted for these differences, outcome at these two levels was comparable, indicating that rather than the health care level providing different treatment quality, patients might with particular characteristics self-select to accessing treatment at specific levels.

In this study, we did not further explore in depth if specific site characteristics, such as characteristics of health care staff, integration of HIV care with other services, or impact of patient load, and patient-health care worker ratio, although many were associated with virological outcome in binary comparison. We instead used the viral load as outcome parameter, that - after adjusting for patient-level differences at treatment start - would then reflect the impact of the overall efficiency and efficacy of the site and all dynamics within the site that hamper or support such an outcome. We found that the viral load cut-off of 400/ml did identify most differences, although the higher cut-off might have been constrained by power and sample size calculations. We thus recommend using this cut-off when assessing the quality of sites and further explore underlying dynamics that may impact site performance.

11.3.3 Patient-Level Factors Affecting Treatment Outcome

In the patient-level analysis against the three different thresholds presented in 10.1.1 main clinical variables associated with treatment failure across all thresholds was a lower absolute CD4 count at study visit, while all other variables were not consistently represented in the models obtained. Overall, our models were able to reproduce associations between virologic failure and patient-level variables that have been described extensively before [96, 103, 104], while other correlations reported elsewhere such as gender [105], cost or time for clinic visit could not be reproduced.

Immunological failure criteria were only associated with virologic failure in VS1000, supporting the decreasing relevance of such surrogate parameters for ART treatment monitoring due to previously reported low sensitivity and specificity to detect the virologic failure and their low clinical benefit over clinical monitoring [106-109]. As focus shifts towards establishing and maintaining a functional “continuum of viral load” [110] the importance of surrogate parameters is diminishing, being replaced by direct treatment monitoring through viral load test.

What can be observed when looking at the models in parallel is that in the VS1000 and VS400 model laboratory parameters associated with inflammation - namely ESR– were significantly associated with virological failure. As infection markers, such correlations with viremia could be explained by temporary viral replication through Co-infections [111-114] independent of drug resistance or as a surrogate parameter for immune activation caused by viral replication itself [115], which has been associated with progression to AIDS [116] and higher overall mortality [117, 118].

In the lower cut-offs, variables describing service satisfaction of the client with the ART service – Building and Waiting time - showed a significant correlation with the outcome, although the events themselves were low and hence we could observe large confidence intervals.

With this in mind, the different models nevertheless can contribute to the discussion about different virologic thresholds: The VS1000 cut-off was associated with other variables that indicate morbidity, and hence might identify treatment failure too late, when it has already had an impact on the immune system. In addition, this cut-off might identify specifically those with an adherence problem rather than those who might have a low-level replication of a resistant virus in the presence of continuous drug intake.

11.3.4 Clinical Score To Predict Virologic Failure And It's Utility In The Public Health Approach.

As outlined in section 6.1.5, the WHO recommended a way to detect virologic failure is either through annually viral load testing in all individuals or targeted viral load testing in individuals suspected of treatment failure. If treatment failure above 1000 copies/ml is detected, viral load should be repeated after intensified adherence counselling and patients should be switched to second-line therapy if no significant drop in viremia can be observed.

Most authors propose to use a score in settings where targeted viral load testing is implemented: the score augments or replaces clinical and immunological criteria in identifying individuals with a risk of virologic failure [81, 163] who are then tested for treatment failure. This scenario is the most common, and the Decision Curve Analysis presented in section 8.9.5. could be used to explore this strategy. As shown, all models had a higher net benefit and thus identified more virologic failure than the clinical or immunological failure criteria at all probability thresholds. The models alone

had further a higher net benefit compared to a combination of a fixed model cut-off, clinical and immunological criteria. The models also showed a higher net benefit compared to a “test all” approach above a 5% probability. Their use would thus maximize the “yield” of viral load testing - which would be the odds of finding a positive test in the ones tested- in settings that are using targeted viral load testing for treatment monitoring. The decision curve also supports a range of possible cut-off points to define if the predicted model probability should be considered “positive for virologic failure” or “negative for virologic failure”. This can be of an advantage in settings where viral load measurements might be infrequently available or the number of available tests varies. If tests are available, the cut-off could be lowered, if tests are sparse, the cut-off could be raised to always ensure that those most at risk will receive the test. A similar approach has been taken by Koller et al.: Using a large dataset including African and Asian routine data, the authors constructed risk charts based on probability models using current and first CD4 count, gender and time on treatment and provided three different cut-offs for initiating viral load tests depending if viral load tests would be available for 10%, 20% or 40% of the population on ART [164].

In respect to performance, the ROC-AUC of all RV288d scores around 0.7, 0.6 and 0.8 in the training, the population and the geographical validation dataset included in the confidence interval ROC-AUC reported by others [81, 165]. The four scores reported by Van Griesven et al. to predict a viral load above 1000 copies/ml, which were simplifications of the prediction score initially developed in Cambodia by Lynen [81], performance in training and validation dataset for the most complex score including laboratory parameters was 0.78 and 0.69 which fell to 0.59 in the validation dataset when the variables were limited to clinical parameters [166].

In comparison with each other, the C-statistic of the three scores developed were similar, with overlapping confidence intervals. In the decision curve analysis, the large score suggested a slightly better net-benefit in a limited range of Threshold Probabilities slight advantage over the other scores the higher risk range.

However, when comparing the score to others reported in the literature, it is important to be aware of other aspects that distinguish our RV288d score from the others reported:

A big strength of our score is that it was developed in a representative sample of the WRSHCP population that included individuals with a large range of treatment duration. Besides the Lynen score [81] and the work of Koll et al. [164], most scores are generated from routine data or prospective cohort studies from single sites where participants have not been recruited as a representative sample of a wider programmatic population. Further, many published scores are not validated, while we performed an intensive validation process using internal validation through bootstrapping as much as external validation taking into account the Limitations discussed in 11.2.2. The performance of this score hence has to be considered representative for the adult program population accessing the WRSHCP more than 6 months on therapy. Other scores were designed to predict treatment outcome at a particular time on ART, commonly one year or 6 months [165, 167, 168]. These scores not aim to identify treatment failure due to resistance development, but rather delayed initial suppression after treatment start due to reduced adherence, to then target these individuals with additional adherence counselling [165]. Our RV288 score developed in this thesis does not exclude the use in such a context, although for this population, variables that refer to treatment start, especially the clinical status at ART start, might be more relevant and might provide better predictors of virologic failure.

The choice of parameters used to predict virologic failure and the feasibility that they can be reliably collected in a standardized fashion in the field further differentiates scores available for viral load prediction:

Most scores use baseline variables collected at treatment start, but do not clearly guide how missing values should be handled in the practical setting to determine failure risk in an individual. For scores like the one developed by Lynen et al. in Cambodia [81] and also validated in an African cohort [72], retrospective values are central and the initial score considered changes from treatment start rather than the values itself, which was the reasons for later simplifications by van Griesen [166]. In all other scores that utilized the immunological WHO failure criteria, a peak CD4 count and a baseline CD4s count are needed to identify the immunological failure. In our study population, such a score would not have been usable for a large number of cases, as often at least one of these variables was missing, prohibiting a complete assessment of immunological failure. While current CD4 count is crucial to predict risk in our score as much as in most others, the inability to establish a CD4 count at

the time of assessment might prevent establishing risk at one point, but not for the remaining duration of the ART treatment. Recognizing that documentation is sometimes limited and patients might transfer from one delivery model to another, switch or re-start ART at different sites, we deliberately did not include variables that had to be collected retrospectively, and the strict limitation to variables that can be collected at the study visit is a further strength of this score. Especially clients with a complex treatment history - for example a woman who started ART in the context of antenatal care and then moved to the CTC after delivery but also anyone switching treatment site – or someone who has been on treatment for longer time can be expected to be more likely to fail treatment and at the same time more likely not have baseline information available. The limitation to current variables allows maximizing the chance that full information is collected in routine clinics and best prediction results are obtained for all.

Scores further vary in complexity of score or underlying variables: The score proposed by Evans et al. includes 14 variables with 5 of them requiring laboratory analysis [163], other scores only rely on one or two variables in addition to or instead of the WHO criteria for immunologic or clinical treatment failure [167]. The complexity of variable generation such as calculating percentage drop of CD4 count of Haemoglobin introduces estimation error in clinical praxis substantial enough to revise the score [166].

Our score placed emphasis on the practicability of the variables collected and should not requires substantial extra time, especially in the patient face-to-face contact to be implemented in the context of a routine visit. While it did include current CD4 count, it does not include clinical diagnosis, which is considered an advantage if the sore should be reproducible or conducted from less-skilled health care workers in the context of a task-shifting approach. However, it should be noted that in a situation where CD4 count is not established through a point of care test but provided through a laboratory, the score will not be completed at the study visit. Further analysis should explore if a retrospective CD4 count could provide similar information and how long the window between score and CD4 count could be to allow for better integration into clinical work. Such further studies would also provide more information on how often and in which intervals this score could be repeated and if it could be an alternative way to monitor treatment over time.

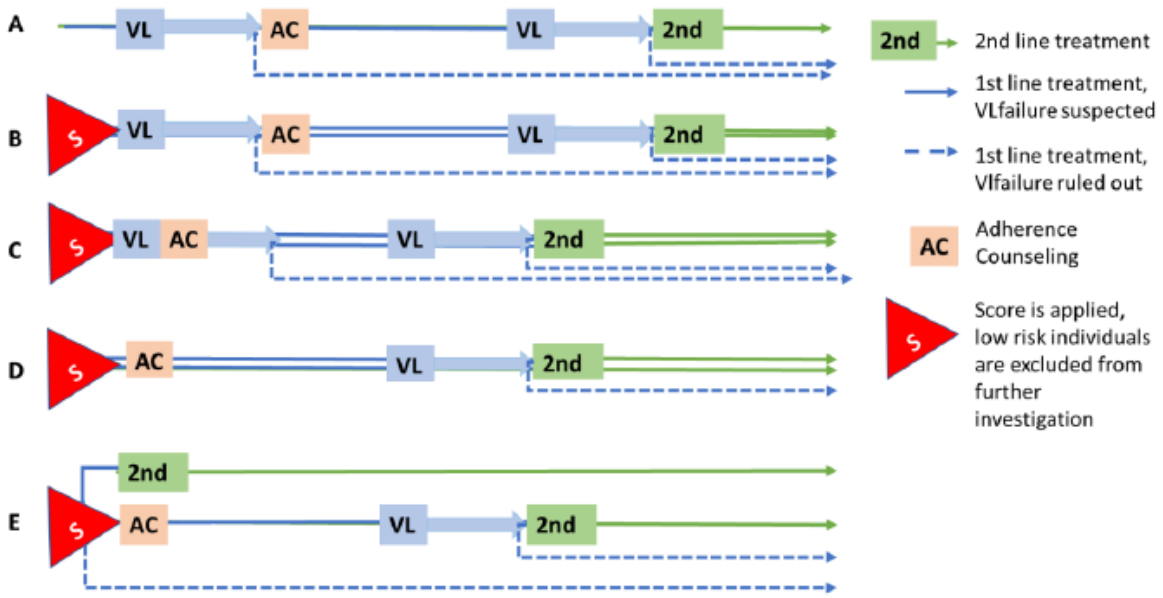
While the scores performed largely comparative, they have all advantages that might make them preferential to each other in particular settings: The score which was used to create the nomogram is, of course, the most practical scoring system as it allows to predict patient risk with a paper-based chart requiring no further equipment and immediate information. The large score, which includes alcohol use and adherence information as additional variables, could be chosen due to its trend to provide slightly higher net benefit and performance which in the current setting did not become statistically significant if the collection of the variables is feasible and an electronic support system would be available to provide the transformation of the variables and calculation of the score. The advantage of the small score, which only included variables that should be available in the patient record and electronic patient files is that this score would not need clinical information from the study visit. Thus, this score would be most useful in case a score should be run as an automated screening tool within a database or even collected to an electronic laboratory system.

Next to using the score to identify individuals that then could undergo viral load testing following the WHO-recommended approach with two consecutive viral loads around an adherence intervention (Figure 32, option B), other use of the score could be considered, which would optimize or change the current WHO guidelines:

As a clinical score provides an immediate result, it could pre-empt laboratory results and trigger adherence counselling [72, 81] allowing the client to move on in the algorithm from the first viral load to adherence counselling while waiting for the viral load result (Figure 32, option C). This would shorten the time between the first and the second viral load test while focussing the resources available to provide quality adherence counselling.

However, the score could also act as the beginning of a triage, that could focus on the use of limited resources on those who would most benefit. Replacing the first viral load, the score with a lower probability threshold could be used to identify those likely to fail, who then could undergo adherence counselling. Following an appropriate period where improved adherence could impact viral suppression, individuals would then undergo a single viral load measurement that would determine virological failure and treatment switch to second line (Figure 32, option D). In modelling studies that simulated a changed WHO algorithm where individuals would be switched to second-line immediately after a first viral load above 1000 copies/ml estimated a median

reduction in AIDS-related morbidity of 31% and mortality of 18% [73] with more individuals with NNRIT resistance mutations being switched to second line. This model did however not account for increase in cost associated with unnecessary switches to second line. Using the score as a digital “first viral load measurement” with adherence counselling would allow reducing the numbers who are above 1000 copies/ml due to adherence challenges and thus decrease the number of individuals switched unnecessarily. While the increase of individuals considered likely to fail would result in an increase in workload for counselling staff, the reduction of viral load test per patients as much as the reduction in costs spent on second line in clients switched unnecessarily are likely to compensate part of the costs.



approach. Only number of individuals tested is increased, **C**: Score used for targeted testing and triggers adherence counselling, reducing time to second line viral load test. **D:** Score replaces 1st viral load test in those at risk of failure and triggers adherence counselling. Only one viral load is needed to identify virologic failure and trigger second line ART start. **E:** Score triages patients by risk, only those with intermediate risk are tested, those with high risk are immediately switched,

Figure 32: Strategic use of a predictive Clinical Score Integrated Into The Procedure To Manage Individuals Failing Antiretroviral Therapy

Finally, a score could streamline procedures and triage individuals in accordance with their triage results: Those with a high risk of failure would directly switch to second-line, those with low risks would not undergo viral load testing and only those with an intermediate risk would receive a viral load test (Figure 32, option E). In this scenario and depending on the cut-offs used to define “high”, “intermediate” and “low risk”, it is

likely that not all individuals with a virological failure would be picked up, but the available resources would be utilized most effectively [81, 164].The digital transformation in the health sector of developing countries with increasing use of electronic patient management systems that collect individual-level data in real-time opens new areas for applications and further developments of predictive scores. If a predictive model would be integrated into patient management or a laboratory software, digital screening for virological failure could be automatized, and patients could be monitored on an ongoing basis, integrating laboratory and clinical results. Integrated software programs could for example link clinical and laboratory data after the client had accessed the clinic and alert health care workers on individuals at risk of treatment failure. The usefulness and efficacy of such clinical decision support systems have been shown before. In a randomized controlled trial in Kenia, a decision support system able to identify immunologic failure and alert health care staff increased clinical action to respond to this immunologic failure by three-fold compared to an unsupported facility [169]. However, digital integration of these tools could transform health care further as score driven processes could range from sending automated re-call SMS to clients with higher risk requesting them to return for further testing, to scheduling adherence counselling at the clinic, or alerting community-based health care workers or lay counsellor to provide adherence support within the community. Clients at risk could be receiving daily reminders to take the medication as part of the adherence counselling or even receive “intelligent drug bottles” for the period between the viral load measurements that could register adherence patterns to inform the interpretation of viral load results. In this context, future work could explore the possibility to use machine learning techniques to improve prediction and provide risk estimations that not only take individual patient risks but also clinic characteristics into account, and allow the model to adapt to changes in the treatment context such as a patient population changing as the proportion of Tanzanians not aware of their positive HIV status is decreasing and more newly diagnosed individuals are enrolled in care.

11.4 Generalisability And Conclusion

Acknowledging the limitations discussed in section 11.2, the results of this analysis are representative for the WRSHCP and findings should be transferable to all individuals accessing care at this program. As the WRSHCP primarily supports the governmental health sector in the regions covered by the study, results are also

generalizable to the national health program in Tanzania and to other countries with a similar setting. However, in the face of a high number of undiagnosed HIV infections in Tanzania, a substantial amount of HIV infected persons in the regions of interest are yet enrolled on ART and thus not represented in the study sample. With intensified focus on reaching those undiagnosed individuals, population characteristics might change and other parameters might become relevant to identify virologic failure.

The clinical score we developed could be shown to perform well across the program including at the site that had been used for the geographical validation. We saw a very low level of model optimism and similar performance in training and validation datasets. The model hence could be used confidently across the sites and clinics of the WRSHCP in Tanzania to predict the risk of failure across the whole adult population on ART. As variables that could describe cultural or other local specifications were not included in the model, it can be expected that the score could be used across Tanzania, and even globally. However, further validation in a population different from the WRSHCP would be important. It is possible that the score would perform well in a comparative setting, as our study sample covered several regions as much as different health care levels. However, with changing populations in care, population characteristics will subsequently change and over time and the performance of this score should be re-evaluated. Similarly, while the score was designed to also be feasible in a community or outreach setting such as a mobile laboratory where it could be deployed through the nomogram and a point of care CD4 counter, our dataset only included facility-based ART provision down to the district health care level. Where ART delivery is taken out of the facility and delivered through community and outreach programs, the population might be different and the predictive performance of the score might be less reliable.

Digitalisation and the availability of electronic routine data allows to generalize not the score itself but rather the procedure of its generation: machine learning algorithms could automate score development based on the theoretical framework and dynamically produce predictive models for specific programs, sites or locations that can identify failure at a specific site without requiring to be equally reliable in other settings.

Such a “site-specific” score would go in line with our finding of the importance of the study site and its direct effect on treatment outcome. Our findings call for a shift in perspective from the discussion on which health care level can provide better care to a more differentiated focus on the role of local dynamics at the treatment sites themselves that are the interface between the health system and the individual. Our results show that sites are impacting treatment outcome presumably through the quality of care they provide, but also that individuals chose the sites they go to. This implies that similar to quality assurance monitoring of service provision in resource-rich countries, programs could employ monitoring of their sites using the methods described in this thesis. Such a monitoring should use the virologic cut-off of 400 copies/ml rather than 1000 copies/ml to have a more sensitive marker of differences and should ensure that assessments control for population-level differences.

12 References

1. (UNDP), U.N.D.P. *Goal 3: Good Health And Wellbeing*. 2015 [cited 2019 12.10.2019]; Available from: https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-3-good-health-and-well-being.html.
2. (NACP), N.A.C.P. *hiv aids in tanzania*. 26.04.2018]; Available from: http://www.nacp.go.tz/site/about/hiv-aids-in-tanzania.
3. Tanzania, M.o.H.C.D.G.E.a.C.M.o. *Summary Sheet: Preliminary Findings - Tanzania HIV Impact Survey (THIS) 2016-2017* 2017 [cited 2018 08/04/2018]; Available from: http://phia.icap.columbia.edu/wp-content/uploads/2017/11/Tanzania_SummarySheet_A4.English.v19.pdf.
4. WHO, *United Republic of Tanzania - Summary Country Profile For HIV/AIDS Treatment Scale-up*. 2005.
5. Tanzania, U.R.O., *HIV/AIDS Care and Treatment Plan 2003-2008*. 2003.
6. Tanzania, T.U.R.o., *The Second National Multi-Sectoral Strategic Framework on HIV and AIDS*. 2007.
7. Tanzania, T.U.R.o., *Tanzania Third National Multi-Sectoral Strategic Framework For HIV and AIDS (2013/14 – 2017/18).* 2013.
8. UNAIDS, *Ending AIDS - Progress towards the 90-90-90 targets*. 2017.
9. UNAIDS. *AIDSinfo | UNAIDS, Data Sheet Tanzania, National, People living with HIV receiving ART (%)*. 2018 [cited 2018 13.04.2018]; Available from: http://aidsinfo.unaids.org/.
10. Dieleman, J.L., et al., *Development assistance for health: past trends, associations, and the future of international financial flows for health.* Lancet, 2016. **387**(10037): p. 2536-44.
11. Schneider, M.T., et al., *Tracking development assistance for HIV/AIDS: the international response to a global epidemic.* Aids, 2016. **30**(9): p. 1475-9.
12. Kieny, M.P., et al., *Strengthening health systems for universal health coverage and sustainable development.* Bull World Health Organ, 2017. **95**(7): p. 537-539.
13. Bain, L.E., C. Nkoke, and J.J.N. Noubiap, *UNAIDS 90–90–90 targets to end the AIDS epidemic by 2020 are not realistic: comment on "Can the UNAIDS 90–90–90 target be achieved? A systematic analysis of national HIV treatment cascades".* BMJ Global Health, 2017. **2**(2).
14. Granich, R., et al., *90-90-90, epidemic control, and ending AIDS: review of global situation and recommendations.* bioRxiv, 2017.
15. Levi, J., et al., *Can the UNAIDS 90-90-90 target be achieved? A systematic analysis of national HIV treatment cascades.* BMJ Glob Health, 2016. **1**(2): p. e000010.
16. (PEPFAR), U.P.s.E.P.F.A.R., *Tanzania - Partnering to Achieve HIV/AIDS Epidemic Control*. 2018.
17. U.S. Military HIV Research Program, W.R.S.H.I.C.P., *PROTOCOL RV 288d - A Virological Assessment of Patients on Antiretroviral Therapy in the US Military HIV Research Program/President's Emergency Plan for AIDS Relief (PEPFAR) – Supported Programs in Africa -Tanzania Site Specific Addendum*. 2011.
18. Phillips, A., et al., *Sustainable HIV Treatment in Africa through Viral Load-Informed Differentiated Care.* Nature, 2015. **528**(7580): p. S68-76.
19. Hoffmann, C.J., J. Maritz, and G.U. Zyl, *CD4 count‐based failure criteria combined with viral load monitoring may trigger worse switch decisions than*

viral load monitoring alone. Tropical Medicine & International Health, 2016. **21**(2): p. 219-223.
20. WHO, *March 2014 supplement to the 2013 consolidated Guidelines on the use of Antiretroviral drugs for treating and preventing HIV infection Recomendation for a Public Health Approach*. 2014.
21. WHO, *Surveillance of HIV Drug Resistance in Adults Recieving ART (Aquired Drug Resistance)* 2014.
22. WHO, *Consolidated ARV guidelines, June 2013*. 2013, Geneva: WHO Press.
23. UNAIDS, *90-90-90 - An ambitious treatment target to help end the AIDS epidemic. .* 2014.
24. Zeh, C., et al., *Evaluation of the performance of Abbott m2000 and Roche COBAS Ampliprep/COBAS Taqman assays for HIV-1 viral load determination using dried blood spots and dried plasma spots in Kenya.* PLoS One, 2017. **12**(6).
25. Palmer, S., et al., *Low-level viremia persists for at least 7 years in patients on suppressive antiretroviral therapy.* Proc Natl Acad Sci U S A, 2008. **105**(10): p. 3879-84.
26. Havlir, D.V., et al., *Prevalence and predictive value of intermittent viremia with combination hiv therapy.* Jama, 2001. **286**(2): p. 171-9.
27. Ryscavage, P., et al., *Significance and clinical management of persistent low-level viremia and very-low-level viremia in HIV-1-infected patients.* Antimicrob Agents Chemother, 2014. **58**(7): p. 3585-98.
28. Mocroft, A., et al., *Is it safe to discontinue primary Pneumocystis jiroveci pneumonia prophylaxis in patients with virologically suppressed HIV infection and a CD4 cell count <200 cells/microL?* Clin Infect Dis, 2010. **51**(5): p. 611-9.
29. Quinn, T.C., et al., *Viral load and heterosexual transmission of human immunodeficiency virus type 1. Rakai Project Study Group.* N Engl J Med, 2000. **342**(13): p. 921-9.
30. Ellman, T.M., et al., *Selecting a viral load threshold for routine monitoring in resource-limited settings: optimizing individual health and population impact.* J Int AIDS Soc, 2017. **20 Suppl 7**.
31. Gray, R.H., et al., *Probability of HIV-1 transmission per coital act in monogamous, heterosexual, HIV-1-discordant couples in Rakai, Uganda.* Lancet, 2001. **357**(9263): p. 1149-53.
32. Ioannidis, J.P., et al., *Perinatal transmission of human immunodeficiency virus type 1 by pregnant women with RNA virus loads <1000 copies/ml.* J Infect Dis, 2001. **183**(4): p. 539-45.
33. Laprise, C., et al., *Virologic failure following persistent low-level viremia in a cohort of HIV-positive patients: results from 12 years of observation.* Clin Infect Dis, 2013. **57**(10): p. 1489-96.
34. Vandenhende, M.A., et al., *Risk of virological failure in HIV-1-infected patients experiencing low-level viraemia under active antiretroviral therapy (ANRS C03 cohort study).* Antivir Ther, 2015. **20**(6): p. 655-60.
35. Taiwo, B., et al., *Antiretroviral drug resistance in HIV-1-infected patients experiencing persistent low-level viremia during first-line therapy.* J Infect Dis, 2011. **204**(4): p. 515-20.
36. Labhardt, N.D., et al., *Should viral load thresholds be lowered?: Revisiting the WHO definition for virologic failure in patients on antiretroviral therapy in resource-limited settings.* Medicine (Baltimore), 2016. **95**(28).
37. Hermans Lucas, M.M.A., Carmona Sergio , Grobbee Diederick , Hofstra Laura Marije, Richman, Tempelman Hugo , Venter Willem D., Wensing Annemarie

Increased Risk of CART Failure after low-leve Viremia under WHO Guidelines, in *Conference on Retrovirus and Opportunistic Infections (CROI)*. 2017 Seattle, Washington.
38. Hermans, *Increased risk of treatment failure after low-level viraemia in a large cohort of South African HIV positive patients treated according to under WHO guidelines*, in *11th INTEREST*. 2017: Lilongwe, Malawi.
39. Samaranayake, A., et al., *Definitions of antiretroviral treatment failure for measuring quality outcomes.* HIV Medicine, 2010. **11**(7): p. 427-431.
40. Programs, R.W.G.H.A. *Clinical Care & Quality Management - Performance Measure Portfolio*. [cited 2019 22.10.2019]; Available from: https://hab.hrsa.gov/clinical-quality-management/performance-measure-portfolio.
41. Rosen, S., M.P. Fox, and C.J. Gill, *Patient retention in antiretroviral therapy programs in sub-Saharan Africa: a systematic review.* PLoS Med, 2007. **4**(10): p. e298.
42. Rosen, S. and M.P. Fox, *Retention in HIV care between testing and treatment in sub-Saharan Africa: a systematic review.* PLoS Med, 2011. **8**(7): p. e1001056.
43. Zurcher, K., et al., *Outcomes of HIV-positive patients lost to follow-up in African treatment programmes.* Trop Med Int Health, 2017. **22**(4): p. 375-387.
44. Joseph Davey, D., et al., *Factors associated with recent unsuppressed viral load in HIV-1-infected patients in care on first-line antiretroviral therapy in South Africa.* Int J STD AIDS, 2018. **29**(6): p. 603-610.
45. Haas, A.D., et al., *Retention and mortality on antiretroviral therapy in sub-Saharan Africa: collaborative analyses of HIV treatment programmes.* J Int AIDS Soc, 2018. **21**(2).
46. Koole, O., et al., *Improved retention of patients starting antiretroviral treatment in Karonga District, northern Malawi, 2005-2012.* J Acquir Immune Defic Syndr, 2014. **67**(1): p. e27-33.
47. Brinkhof, M.W., et al., *Adjusting mortality for loss to follow-up: analysis of five ART programmes in sub-Saharan Africa.* PLoS One, 2010. **5**(11): p. e14149.
48. Brinkhof, M.W., M. Pujades-Rodriguez, and M. Egger, *Mortality of patients lost to follow-up in antiretroviral treatment programmes in resource-limited settings: systematic review and meta-analysis.* PLoS One, 2009. **4**(6): p. e5790.
49. Somi, G., et al., *Low mortality risk but high loss to follow-up among patients in the Tanzanian national HIV care and treatment programme.* Trop Med Int Health, 2012. **17**(4): p. 497-506.
50. Lamb, M.R., et al., *High attrition before and after ART initiation among youth (15-24 years of age) enrolled in HIV care.* Aids, 2014. **28**(4): p. 559-68.
51. Koole, O., et al., *Retention and risk factors for attrition among adults in antiretroviral treatment programmes in Tanzania, Uganda and Zambia.* Trop Med Int Health, 2014. **19**(12): p. 1397-410.
52. (CDC), U.C.f.D.C., *Differences between HIV-Infected men and women in antiretroviral therapy outcomes - six African countries, 2004-2012.* MMWR Morb Mortal Wkly Rep, 2013. **62**(47): p. 945-52.
53. Crawford, K., et al., *Evaluation of treatment outcomes for patients on first-line regimens in US President's Emergency Plan for AIDS Relief (PEPFAR) clinics in Uganda: predictors of virological and immunological response from RV288 analyses.* HIV Med, 2014.

54. Thielman, N.M., et al., *Reduced Adherence to Antiretroviral Therapy among HIV-infected Tanzanians Seeking Cure from the Loliondo Healer.* J Acquir Immune Defic Syndr, 2014. **65**(3): p. e104-9.
55. Wanyama, J.N., et al., *Persons living with HIV infection on antiretroviral therapy also consulting traditional healers: a study in three African countries.* Int J STD AIDS, 2017. **28**(10): p. 1018-1027.
56. Al-Dakkak, I., et al., *The impact of specific HIV treatment-related adverse events on adherence to antiretroviral therapy: a systematic review and meta-analysis.* AIDS Care, 2013. **25**(4): p. 400-14.
57. Ferguson, L., et al., *Linking women who test HIV-positive in pregnancy-related services to HIV care and treatment services in Kenya: a mixed methods prospective cohort study.* PLoS One, 2014. **9**(3): p. e89764.
58. Nachega, J.B., et al., *Association between antiretroviral therapy adherence and employment status: systematic review and meta-analysis.* Bull World Health Organ, 2015. **93**(1): p. 29-41.
59. Chen, W.T., et al., *Engagement with Health Care Providers Affects Self-Efficacy, Self-Esteem, Medication Adherence and Quality of Life in People Living with HIV.* J AIDS Clin Res, 2013. **4**(11): p. 256.
60. Daigle, G.T., et al., *System-level factors as predictors of adherence to clinical appointment schedules in antiretroviral therapy in Cambodia.* AIDS Care, 2015. **27**(7): p. 836-43.
61. Chaiyachati, K.H., et al., *Interventions to improve adherence to antiretroviral therapy: a rapid systematic review.* Aids, 2014. **28 Suppl 2**: p. S187-204.
62. Adjorlolo-Johnson, G., et al., *Scaling up pediatric HIV care and treatment in Africa: clinical site characteristics associated with favorable service utilization.* J Acquir Immune Defic Syndr, 2013. **62**(1): p. e7-e13.
63. Rackal, J.M., et al., *Provider training and experience for people living with HIV/AIDS.* Cochrane Database Syst Rev, 2011(6): p. Cd003938.
64. Solomons, D.J., et al., *Factors influencing the confidence and knowledge of nurses prescribing antiretroviral treatment in a rural and urban district in the Western Cape province.* South Afr J HIV Med, 2019. **20**(1): p. 923.
65. Kredo, T., et al., *Decentralising HIV treatment in lower- and middle-income countries.* Cochrane Database Syst Rev, 2013(6): p. Cd009987.
66. Di Giorgio, L., et al., *The potential to expand antiretroviral therapy by improving health facility efficiency: evidence from Kenya, Uganda, and Zambia.* BMC Med, 2016. **14**(1): p. 108.
67. Colvin, C.J., et al., *A systematic review of health system barriers and enablers for antiretroviral therapy (ART) for HIV-infected pregnant and postpartum women.* PLoS One, 2014. **9**(10): p. e108150.
68. van Lettow, M., et al., *Towards elimination of mother-to-child transmission of HIV: performance of different models of care for initiating lifelong antiretroviral therapy for pregnant women in Malawi (Option B+).* J Int AIDS Soc, 2014. **17**: p. 18994.
69. Handford, C.D., et al., *Organization of care for persons with HIV-infection: a systematic review.* AIDS Care, 2017. **29**(7): p. 807-816.
70. Hornberger, J., et al., *A systematic review of cost-utility analyses in HIV/AIDS: implications for public policy.* Med Decis Making, 2007. **27**(6): p. 789-821.
71. WHO, *What's New In Treatment Monitroing: Viral Laod And Cd4Testing - Update July 2017*, in *HIV Treatment And Care Information Note*, W.H. Organisation, Editor. 2017.

72. Labhardt, N.D., et al., *A Clinical Prediction Score in Addition to WHO Criteria for Anti-Retroviral Treatment Failure in Resource-Limited Settings - Experience from Lesotho.* PLOS ONE, 2012. **7**(10): p. e47937.
73. Shroufi, A., et al., *Simplifying switch to second-line antiretroviral therapy in sub Saharan Africa: predicted effect of using a single viral load to define efavirenz-based first-line failure.* AIDS (London, England), 2019. **33**(10): p. 1635-1644.
74. Zhang, Z., H. Zhang, and M.K. Khanal, *Development of scoring system for risk stratification in clinical medicine: a step-by-step tutorial.* Ann Transl Med, 2017. **5**(21): p. 436.
75. Wynants, L., et al., *Untapped potential of multicenter studies: a review of cardiovascular risk prediction models revealed inappropriate analyses and wide variation in reporting.* Diagn Progn Res, 2019. **3**: p. 6.
76. Steyerberg, E.W. and Y. Vergouwe, *Towards better clinical prediction models: seven steps for development and an ABCD for validation.* European Heart Journal, 2014. **35**(29): p. 1925-1931.
77. Kahle, E.M., et al., *An empiric risk scoring tool for identifying high-risk heterosexual HIV-1-serodiscordant couples for targeted HIV-1 prevention.* J Acquir Immune Defic Syndr, 2013. **62**(3): p. 339-47.
78. Pintye, J., et al., *A Risk Assessment Tool for Identifying Pregnant and Postpartum Women Who May Benefit From Preexposure Prophylaxis.* Clin Infect Dis, 2017. **64**(6): p. 751-758.
79. Hanifa, Y., et al., *A clinical scoring system to prioritise investigation for tuberculosis among adults attending HIV clinics in South Africa.* PLOS ONE, 2017. **12**(8): p. e0181519.
80. Abouyannis, M., et al., *Development and validation of systems for rational use of viral load testing in adults receiving first-line ART in sub-Saharan Africa.* Aids, 2011. **25**(13): p. 1627-35.
81. Lynen, L., et al., *An algorithm to optimize viral load testing in HIV-positive patients with suspected first-line antiretroviral therapy failure in Cambodia.* J Acquir Immune Defic Syndr, 2009. **52**(1): p. 40-8.
82. Austin, P.C., *An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies.* Multivariate Behav Res, 2011. **46**(3): p. 399-424.
83. Austin, P.C., *The performance of different propensity-score methods for estimating differences in proportions (risk differences or absolute risk reductions) in observational studies.* Stat Med, 2010. **29**(20): p. 2137-48.
84. Austin, P.C., N. Jembere, and M. Chiu, *Propensity score matching and complex surveys.* Stat Methods Med Res, 2018. **27**(4): p. 1240-1257.
85. Linden, A., *Combining propensity score-based stratification and weighting to improve causal inference in the evaluation of health care interventions.* J Eval Clin Pract, 2014. **20**(6): p. 1065-71.
86. Rubin, D.B., *The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials.* Statistics in Medicine, 2007. **26**(1): p. 20-36.
87. Rosenbaum Paul R., R.D.B., *The Central Role of the Propensity Score in Observational Studies for Causal Effect.* Biometrika, 1983. **70**(1): p. 41-55.
88. Spreeuwenberg, M.D., et al., *The multiple propensity score as control for bias in the comparison of more than two treatment arms: an introduction from a case study in mental health.* Med Care, 2010. **48**(2): p. 166-74.
89. The United Republic of Tanzania, M.o.H.a.S.W., Tanzania Mainland, *National Guidelines for the Management of HIV and AIDS*. 4th ed, ed. N.A.C.P. (NACP).

January 2012, Dar Es Salaam: Ministry of Health and Social Welfare, Tanzania, National AIDS Control Programme (NACP).
90. WHO, *Case Definitions of HIV for Surveillance and Revised Clinical Staging and Immunological Classification of HIV- related Diseases in Adults and Children* 2007.
91. *<2013 WHO Guidelines.pdf>.*
92. Moons, K.G., et al., *Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration.* Ann Intern Med, 2015. **162**(1): p. W1-73.
93. Pavlou, M., et al., *How to develop a more accurate risk prediction model when there are few events.* BMJ : British Medical Journal, 2015. **351**: p. h3868.
94. WHO. *Health Topics - Body mass index - BMI*. 2019 [cited 2019 11.03.]; Available from: http://www.euro.who.int/en/health-topics/disease-prevention/nutrition/a-healthy-lifestyle/body-mass-index-bmi.
95. Saathoff, E., et al., *Laboratory reference values for healthy adults from southern Tanzania.* Trop Med Int Health, 2008. **13**(5): p. 612-25.
96. StataCorp., *Stata Statistical Software: Release 13. College Station, TX: StataCorp LP. .* 2013.
97. McCaffrey, D.F., et al., *A tutorial on propensity score estimation for multiple treatments using generalized boosted models.* Stat Med, 2013. **32**(19): p. 3388-414.
98. Cepeda, M.S., et al., *Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders.* Am J Epidemiol, 2003. **158**(3): p. 280-7.
99. Greenland, S., R. Daniel, and N. Pearce, *Outcome modelling strategies in epidemiology: traditional methods and basic alternatives.* Int J Epidemiol, 2016. **45**(2): p. 565-75.
100. Rosenbaum, P.R. and D.B. Rubin, *Reducing Bias in Observational Studies Using Subclassification on the Propensity Score.* Journal of the American Statistical Association, 1984. **79**(387): p. 516-524.
101. Robins, J.M. and S. Greenland, *The role of model selection in causal inference from nonexperimental data.* Am J Epidemiol, 1986. **123**(3): p. 392-402.
102. Cefalu M., B.M., *Propensity Scores for Multipel Treatments - A Tutorial on the MNPS Command for Stata Users*. 2017, RAND Corporation, : Santa Monica,, Calif.
103. Greg Ridgeway, D.M., Andrew Morral, Beth Ann Griffin, Lane Burgette, *twang: Toolkit for Weighting and Analysis of Nonequivalent Groups*. 2017.
104. Cooperation, R., *Stata Macros - MNPS Command for Stata Users*.
105. Linden, A. and J.L. Adams, *Using propensity score-based weighting in the evaluation of health management programme effectiveness.* Journal of Evaluation in Clinical Practice, 2010. **16**(1): p. 175-179.
106. Linden, A. and J.L. Adams, *Combining the regression discontinuity design and propensity score-based weighting to improve causal inference in program evaluation.* J Eval Clin Pract, 2012. **18**(2): p. 317-25.
107. StataCorp., *Survey Data Reference Manual Release 15*. 2017.
108. Bergstra, S.A., et al., *Three handy tips and a practical guide to improve your propensity score models.* RMD Open, 2019. **5**(1): p. e000953.
109. Breiman, L., *Classification and Regression Trees*. 1984, New York: Routledge.
110. Cohen, J., *Statistical Power Analysis for the Behavioral Sciences*. 1988, New York: Routledge.

111. Lee, Y.H., H. Bang, and D.J. Kim, *How to Establish Clinical Prediction Models.* Endocrinol Metab (Seoul), 2016. **31**(1): p. 38-44.
112. Collins, G.S., et al., *Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement.* Bmj, 2015. **350**: p. g7594.
113. StataCorp, *Stata Statistical Software*, in *Release 16*. 2019, StataCorp LLC.: College Station, TX.
114. Harrell, F.E., Jr., et al., *Regression models for prognostic prediction: advantages, problems, and suggested solutions.* Cancer Treat Rep, 1985. **69**(10): p. 1071-77.
115. Pace, N.L., L.H. Eberhart, and P.R. Kranke, *Quantifying prognosis with risk predictions.* Eur J Anaesthesiol, 2012. **29**(1): p. 7-16.
116. Harrell, F.E., Jr., K.L. Lee, and D.B. Mark, *Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors.* Stat Med, 1996. **15**(4): p. 361-87.
117. *Population Health methods - Risk Prediction*. [cited 08.12.2019; Available from: https://www.mailman.columbia.edu/research/population-health-methods/risk-prediction.
118. Austin, P.C. and E.W. Steyerberg, *Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers.* Stat Med, 2014. **33**(3): p. 517-35.
119. Bouwmeester, W., et al., *Internal validation of risk models in clustered data: a comparison of bootstrap schemes.* Am J Epidemiol, 2013. **177**(11): p. 1209-17.
120. Vickers, A.J., B. Van Calster, and E.W. Steyerberg, *Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests.* Bmj, 2016. **352**: p. i6.
121. Vickers, A.J. and E.B. Elkin, *Decision curve analysis: a novel method for evaluating prediction models.* Med Decis Making, 2006. **26**(6): p. 565-74.
122. Vickers, A.J. and A.M. Cronin, *Traditional statistical methods for evaluating prediction models are uninformative as to clinical value: towards a decision analytic framework.* Semin Oncol, 2010. **37**(1): p. 31-8.
123. Steyerberg, E.W. and A.J. Vickers, *Decision curve analysis: a discussion.* Med Decis Making, 2008. **28**(1): p. 146-9.
124. Vickers, A.J., B. van Calster, and E.W. Steyerberg, *A simple, step-by-step guide to interpreting decision curve analysis.* Diagnostic and Prognostic Research, 2019. **3**(1): p. 18.
125. Van Calster, B. and A.J. Vickers, *Calibration of Risk Prediction Models:Impact on Decision-Analytic Performance.* Medical Decision Making, 2015. **35**(2): p. 162-169.
126. Kerr, K.F., et al., *Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use.* J Clin Oncol, 2016. **34**(21): p. 2534-40.
127. Steyerberg, E.W., et al., *Assessing the performance of prediction models: a framework for traditional and novel measures.* Epidemiology, 2010. **21**(1): p. 128-38.
128. StataCorpLP, *Survey 13 Documentation*. Stata Press.
129. Dalhatu, I., et al., *Outcomes of Nigeria's HIV/AIDS Treatment Program for Patients Initiated on Antiretroviral Treatment between 2004-2012.* PLoS One, 2016. **11**(11): p. e0165528.

130. Roberto, G.D.D.M. *Citing references for Stata's cluster-correlated robust variance estimates* 06.02.2019]; Available from: https://www.stata.com/support/faqs/statistics/references/.
131. Rogers, W.H., *Regression standard errors in clustered samples.* Stata Technical Bulletin, 1993.(13): p. 19–23.
132. Williams, R.L., *A note on robust variance estimation for cluster-correlated data.* Biometrics 2000. **56**: p. 645–646.
133. Wooldridge, J.M., *Econometric Analysis of Cross Section and Panel Data*. 2002, Cambridge MA: MIT Press.
134. Froot, K.A., *Consistent covariance matrix estimation with cross-sectional dependence and heteroskedasticity in financial data.* Journal of Financial and Quantitative Analysis, 1989. **24**: p. 333–355.
135. Bouwmeester, W., et al., *Prediction models for clustered data: comparison of a random intercept and standard regression model.* BMC Med Res Methodol, 2013. **13**: p. 19.
136. Ni, H., et al., *Prediction models for clustered data with informative priors for the random effects: a simulation study.* BMC Med Res Methodol, 2018. **18**(1): p. 83.
137. Bell, A. and K. Jones, *Explaining Fixed Effects: Random Effects Modeling of Time-Series Cross-Sectional and Panel Data.* Political Science Research and Methods, 2014. **3**(1): p. 133-153.
138. Carle, A.C., *Fitting multilevel models in complex survey data with design weights: Recommendations.* BMC Medical Research Methodology, 2009. **9**(1): p. 49.
139. Anthony G. Turner, G.A., Amy O. Tsui, Marilyn Wilkinson, Robert Magnani, *Sampling Manual for Facility Surveys - For Population, Maternal Health, Child Health and STD Programs in Developing Countries*. July 2001, MEASURE Evaluation, USAID.
140. Lee, E.S. and R.N. Forthofer, *Analyzing Complex Survey Data*. 2005: SAGE Publications.
141. Hawkins, C., et al., *HIV virological failure and drug resistance in a cohort of Tanzanian HIV-infected adults.* J Antimicrob Chemother, 2016. **71**(7): p. 1966-74.
142. Project, P., *Tanzania HIV Impact Survey (THIS) 2016-2017- Summary Sheet: Preliminary Findings.* 2017.
143. McMahon, J.H., et al., *Viral suppression after 12 months of antiretroviral therapy in low- and middle-income countries: a systematic review.* Bull World Health Organ, 2013. **91**(5): p. 377-385e.
144. Fatti, G., A. Grimwood, and P. Bock, *Better antiretroviral therapy outcomes at primary healthcare facilities: an evaluation of three tiers of ART services in four South African provinces.* PLoS One, 2010. **5**(9): p. e12888.
145. Crawford, K.W., et al., *East Meets West: A Description of HIV-1 Drug Resistance Mutation Patterns of Patients Failing First Line Therapy in PEPFAR Clinics from Uganda and Nigeria.* AIDS Res Hum Retroviruses, 2014.
146. Khumalo, P.G., Y.J. Chou, and C. Pu, *Antiretroviral treatment attrition in Swaziland: a population-based study.* Epidemiol Infect, 2016. **144**(16): p. 3474-3482.
147. Bilinski, A., et al., *Distance to care, enrollment and loss to follow-up of HIV patients during decentralization of antiretroviral therapy in Neno District, Malawi: A retrospective cohort study.* PLoS One, 2017. **12**(10): p. e0185699.

148. Terzian, A.S., et al., *Identifying Spatial Variation Along the HIV Care Continuum: The Role of Distance to Care on Retention and Viral Suppression.* AIDS and behavior, 2018. **22**(9): p. 3009-3023.
149. Tanzania Commission for AIDS (TACAIDS), Z.A.C.Z., *Tanzania HIV Impact Survey (THIS) 2016-2017: FinalReport*. 2018: Dar es Salaam, Tanzania.
150. Liwanag, H.J. and K. Wyss, *What conditions enable decentralization to improve the health system? Qualitative analysis of perspectives on decision space after 25 years of devolution in the Philippines.* PLoS One, 2018. **13**(11): p. e0206809.
151. Okonkwo, P., et al., *Treatment outcomes in a decentralized antiretroviral therapy program: a comparison of two levels of care in north central Nigeria.* AIDS Res Treat, 2014. **2014**: p. 560623.
152. Nachega, J.B., et al., *Community-Based Interventions to Improve and Sustain Antiretroviral Therapy Adherence, Retention in HIV Care and Clinical Outcomes in Low- and Middle-Income Countries for Achieving the UNAIDS 90-90-90 Targets.* Curr HIV/AIDS Rep, 2016. **13**(5): p. 241-55.
153. Lazarus, D.D., et al., *The lack of an effect by insulin or insulin-like growth factor-1 in attenuating colon-2-mediated cancer cachexia.* Cancer Lett, 1996. **103**(1): p. 71-7.
154. Ciapponi, A., et al., *Delivery arrangements for health systems in low-income countries: an overview of systematic reviews.* The Cochrane database of systematic reviews, 2017. **9**(9): p. CD011083-CD011083.
155. Tsui, S., et al., *Identifying models of HIV care and treatment service delivery in Tanzania, Uganda, and Zambia using cluster analysis and Delphi survey.* BMC Health Serv Res, 2017. **17**(1): p. 811.
156. Haberer, J.E., et al., *Improving antiretroviral therapy adherence in resource-limited settings at scale: a discussion of interventions and recommendations.* J Int AIDS Soc, 2017. **20**(1): p. 21371.
157. Govindasamy, D., et al., *Interventions to improve or facilitate linkage to or retention in pre-ART (HIV) care and initiation of ART in low- and middle-income settings--a systematic review.* J Int AIDS Soc, 2014. **17**: p. 19032.
158. Penn, A.W., et al., *Supportive interventions to improve retention on ART in people with HIV in low- and middle-income countries: A systematic review.* PLoS One, 2018. **13**(12): p. e0208814.
159. Kendall, C.E., et al., *A population-based study comparing patterns of care delivery on the quality of care for persons living with HIV in Ontario.* BMJ Open, 2015. **5**(5): p. e007428.
160. Landovitz, R.J., et al., *Quality of Care for HIV/AIDS and for Primary Prevention by HIV Specialists and Nonspecialists.* AIDS Patient Care STDS, 2016. **30**(9): p. 395-408.
161. Agaba, P.A., et al., *Retention in Differentiated Care: Multiple Measures Analysis for a Decentralized HIV Care and Treatment Program in North Central Nigeria.* J AIDS Clin Res, 2018. **9**(2).
162. Burua, A., F. Nuwaha, and P. Waiswa, *Adherence to standards of quality HIV/AIDS care and antiretroviral therapy in the West Nile Region of Uganda.* BMC health services research, 2014. **14**: p. 521-521.
163. Evans, D.H., et al., *CD4 criteria improves the sensitivity of a clinical algorithm developed to identify viral failure in HIV-positive patients on antiretroviral therapy.* J Int AIDS Soc, 2014. **17**: p. 19139.
164. Koller, M., et al., *Implementation and Operational Research: Risk Charts to Guide Targeted HIV-1 Viral Load Monitoring of ART: Development and*

Validation in Patients From Resource-Limited Settings. Journal of acquired immune deficiency syndromes (1999), 2015. **70**(3): p. e110-e119.
165. Mbengue, M.A.S., et al., *Clinical predictor score to identify patients at risk of poor viral load suppression at six months on antiretroviral therapy: results from a prospective cohort study in Johannesburg, South Africa.* Clinical epidemiology, 2019. **11**: p. 359-373.
166. van Griensven, J., et al., *Simplified Clinical Prediction Scores to Target Viral Load Testing in Adults with Suspected First Line Treatment Failure in Phnom Penh, Cambodia.* PLOS ONE, 2014. **9**(2): p. e87879.
167. Reepalu, A., et al., *Development of an algorithm for determination of the likelihood of virological failure in HIV-positive adults receiving antiretroviral therapy in decentralized care.* Global health action, 2017. **10**(1): p. 1371961-1371961.
168. Robbins, G.K., et al., *Predicting virologic failure in an HIV clinic.* Clin Infect Dis, 2010. **50**(5): p. 779-86.
169. Oluoch, T., et al., *Effect of a clinical decision support system on early action on immunological treatment failure in patients with HIV in Kenya: a cluster randomised controlled trial.* The Lancet HIV, 2016. **3**(2): p. e76-e84.

13 Annexe

13.1 List Of Publications

Huber J, Bauer D, Hoelscher M, Kapungu J, Kroidl A, Lennemann T, Maganga L, Opitz O, Salehe O, Sigauke A, Fischer MR, Kiessling C., Evaluation of health research capacity strengthening trainings on individual level: validation of a questionnaire J Eval Clin Pract. 2014 Aug;20(4):390-5. doi: 10.1111/jep.12143. Epub 2014 May 15

Biru T, Lennemann T, Stürmer M, Stephan C, Nisius G, Cinatl J, Staszewski S, Gürtler LG., Human immunodeficiency virus type-1 group M quasispecies evolution: diversity and divergence in patients co-infected with active tuberculosis. Med Microbiol Immunol. 2010 Aug 10. [Epub ahead of print]

Rabenau, Lennemann (Shared first autors), Kircher, Gürtler, Staszewski, Preiser, McPherson, Allwin, Doerr, Prevalence and gender specific immune response to opportunistic infections in HIV-infected patients in Lesotho, Sex Transm Dis. 2010 Jul;37(7):454-9.

13.2 Statement On Pre-Release And Contribution

I, Tessa Lennemann, hereby declare that

I have not previously published or submitted for publication any of this research. My role in this program has been the Principal Investigator in RV288, ensuring the protocol conform implementation of the study. I further have done all analysis and writing in this thesis.

Tessa Lennemann

13.3 Acknowledgment

I would like to acknowledge the RV 288 team for implementing the study and the participants for consenting to it. I would further like to acknowledge and thank my husband who never gave up on me.

13.4 Supplemental Material

13.4.1 Overlap Plots Of The GBM Models

The assumption of overlap requires that all units of the dataset have a non-zero probability to be included in other groups than the one they are in and that this non-zero probability is the propensity score. In GBM models with more than two exposure groups, probabilities are calculated for each of the exposures against the pooled group of others not exposed. The overlap plots present these probabilities for each of the exposure groups

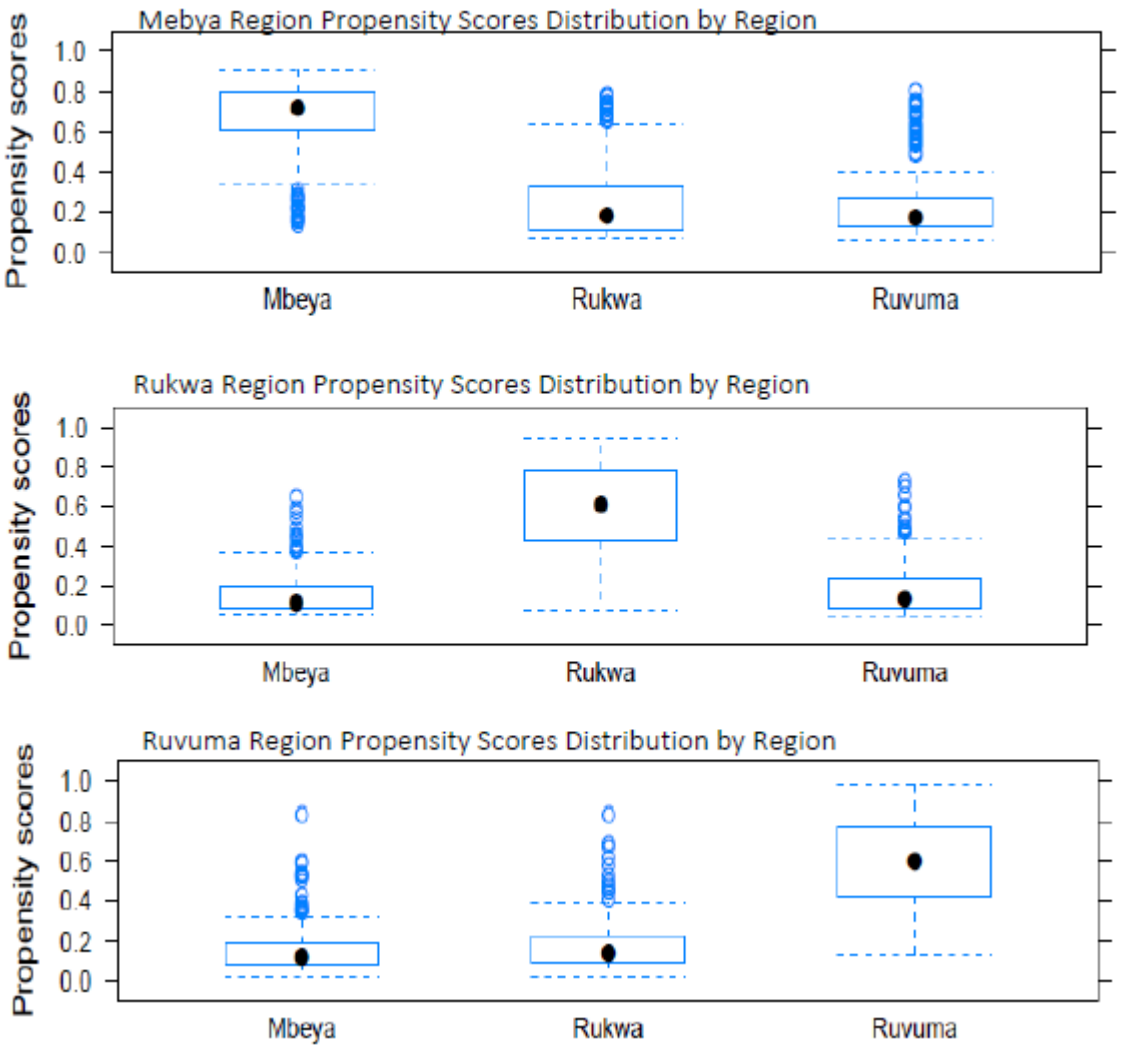


Figure 34: Distribution of Regional Propensity Scores

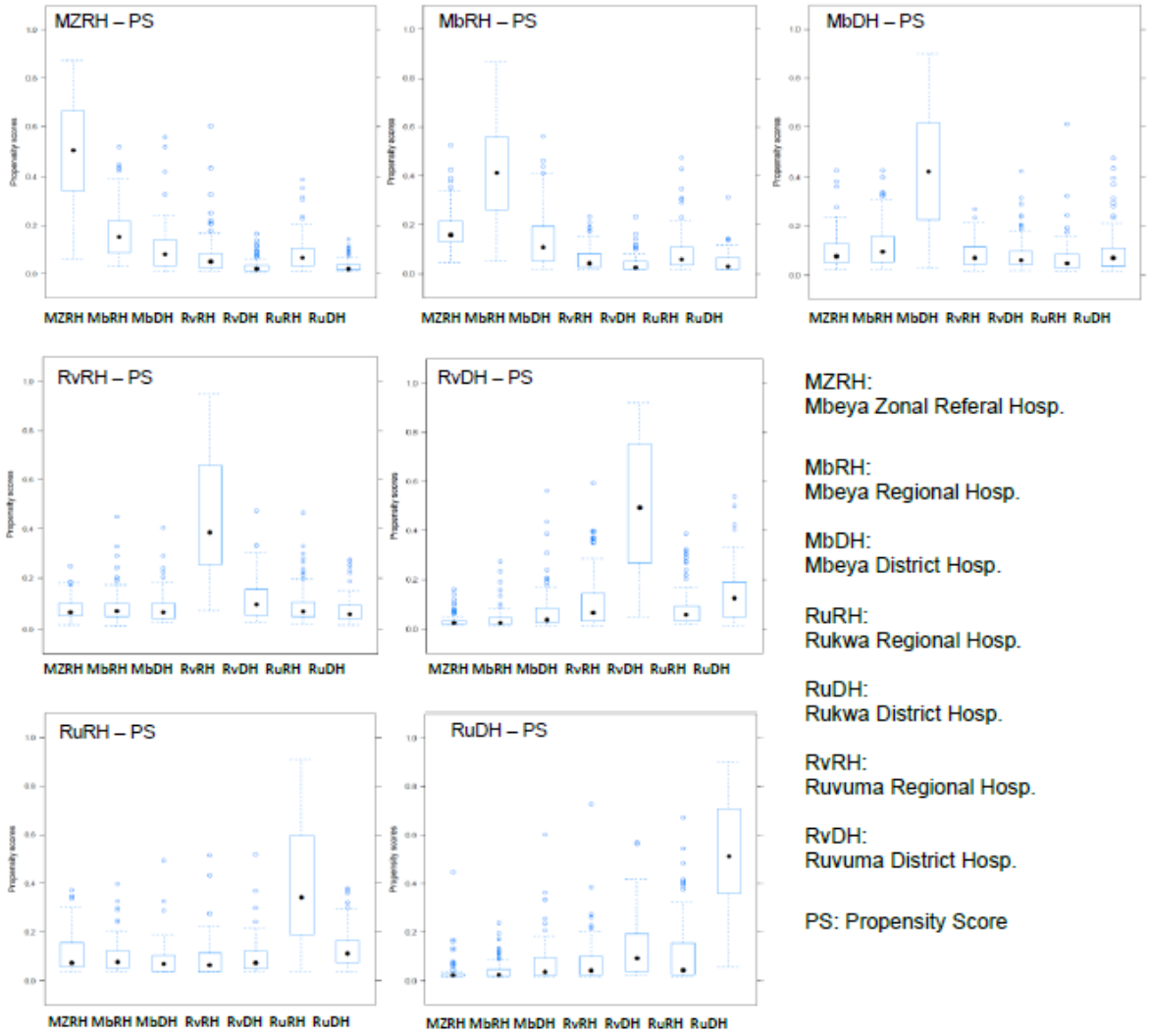


Figure 35: Distribution of Site Propensity Scores

13.4.2 Table 8: Level 1 Variables Assessed For Inclusion In Predictive Model, Association With Outcome In The Training Dataset And Rationale For Inclusion

Variable Name	% missing in full RV288 datatset	Total number of events in RV288 study population	Association with outcome (VS400) in training datatset	Paired with Step in Causal Pathway	Rank	Rationale for Rank
Mean Copuscular Volumen (MCV)	1	523	0.578	D0 Active drug	3	MCV could be a good marker for adherence to AZT continaing regimen, but AZT is phased out of first line therapy so varibale can be expected to have less relevance in the future
Time on ART	0	NA	0.099	D1 ARV Resistance	1	Exposure time to ART is one of the main factors for resistance development and subsequent treatment failure
Ever had TB/ART co-medication	0	62	0.222	D1 ARV Resistance	2	Very relevant but few cases
ALT at SV	0	7	0.162	D1 Metabolism/Resorption	4	no association with outcome aligns with very indirect causal relationship between ALT and viral load
Creatinin at SV	0	16	0.966	D1 Metabolism/Resorption	4	no association with V400, TDF major confounder, many missing values in study sample, not part of routine Labpanel for all clients
Ever had liver disease	37	5	0.829	D1 Metabolism/Resorption	4	not enough cases
Ever had renal disease	37	8	0.762	D1 Metabolism/Resorption	4	not enough cases
ALT at BL	45	18	0.661	D1 Metabolism/Resorption	4	not enough cases
Lypodrystrophy at SV	0	20	0.582	D1 Side effects	4	not enough cases
Polyneuropathy at SV	0	26	0.976	D1 Side effects	4	not enough cases
Number of tablets per day	0	NA	0.836	D2 Regular ART Intake	2	Pill burden known to be associated with adherence
Dosage frequency/per day	0	NA	0.210	D2 Regular ART Intake	4	not enough variance as 93% of population take BD.
Number of ARVs per day	0	NA	0.229	D2 Regular ART Intake	4	would change with available medication and full pill burden (tabsday) should be the better measure

Drinks any alcohol	0	149	0.002	D3 CNS	1	Association with viral failure, easy screenign question for field
Ever had neurologic diseases	0	4	0.530	D3 CNS	4	not enough cases
Reason not to take ARV - Depression	0	10	0.275	D3 CNS	4	not enough cases
Ever had depression	37	32	0.309	D3 CNS	4	not enough cases
Ever had migrane	37	55	0.320	D3 CNS	4	not enough cases
Takes any co-medications including prophylaxis next to ART	0	313	0.206	D3 Co-Medication	2	Simplified to number of CO-medication
Any prophylaxis	0	264	0.141	D3 Co-Medication	2	Prof not directly influencing VL, but could identify particular patient group who are still of risk of opportunitstic infections but at the same time enaged in care.
Any therapeutic medication next to ART, excluding prophylaxis	0	63	0.851	D3 Co-Medication	4	CO-medication could be a surrogate for compromised clinical disease status and risk of interaction with ART, but not enough cases
Current dabetic medication	0	4	0.024	D3 Co-Medication	4	not enough cases
Current hypertensive therapy	0	40	0.585	D3 Co-Medication	4	not enough cases
Current pulmonal medication	0	3	0.001	D3 Co-Medication	4	not enough cases
Number of reasons to miss ART	0	125	0.429	D3 Disease Perception	2	Might present a grade of difficulty to adhere to ART
Number of reasons to miss study visit	0	84	0.692	D3 Disease Perception	2	Might present a grade of difficulty to access clinic
Uses traditional healthcare next to CTC	0	85	0.970	D3 Disease Perception	2	Could indicate particular adherence pattern and risk of interaction with ART
Reaons to miss ART - share medication	0	1	0.657	D3 Disease Perception	4	not enough cases
Reason to miss ART - ill health	0	5	0.428	D3 Disease Perception	4	not enough cases
Reason to miss ART- feeling better	0	6	0.069	D3 Disease Perception	4	not enough cases

Reason to miss ART-forgetting	0	71	0.241	D3 Disease Perception	4	not enough cases
Reason to miss ART-toxicity	0	5/10	0.031	D3 Disease Perception	4	not enough cases
Uses traditional healers	0	57	0.718	D3 Disease Perception	4	not enough cases
Uses traditional remedies	0	76	0.549	D3 Disease Perception	4	not enough cases
Missed any follow up visit	0	27	0.030	D3 Regular Pick up visits	1	Rough measure of adherence, easy to collect
Number of missed follow up visits	0	27	0.028	D3 Regular Pick up visits	2	Introduces scale, but very little variance in study sample
Reason to miss visits - stigma and discrimination	0	7	0.152	D3 Regular Pick up visits	4	not enough cases
Eliglibililty criteria for treatment start	0	NA	0.699	D4 Pt /Clinic Interface	2	this variable has changed with new guideline test and treat
Use of EFV or NVP as NNRTI backbone	0	NA	0.350	D4 Pt /Clinic Interface	2	Relevance according to literature but no association in study sample
Time from eliglibility to treatment start	2	NA	0.203	D4 Pt /Clinic Interface	2	technically could be used but might have changed due to programmatic change
Time between diagnosis and access to care	6	NA	0.278	D4 Pt /Clinic Interface	2	Relevanc according to literature but no association in study sample
Service Satisfaction - long waiting time	0	20	0.415	D4 Pt /Clinic Interface	2	no association to outcome in study sample but in literature, and low cases
Duration between CD4 count and ART start	9	NA	0.007	D4 Pt /Clinic Interface	3	This variable is ambigous, as time between ART start and CD4 count can be influenced by various effects within the clinic
Way adherence counceling is provided	0	NA		D4 Pt /Clinic Interface	3	not much variability - either only group counseling or combination
Functional status at treatment start	0	NA	0.353	D4 Pt /Clinic Interface	4	very little variance
NRTI backbone in current regimen	0	649/53	0.139	D4 Pt /Clinic Interface	4	little variance
Reason not to take ARV - Pharamcy out of stock	0	4/1	0.662	D4 Pt /Clinic Interface	4	not enough cases
Reason not to take ARV - has no pills	0	1	0.642	D4 Pt /Clinic Interface	4	not enough cases

Time from entry to care to eliglibility	3	NA	0.274	D4 Pt /Clinic Interface	4	this variable has changed with new guideline test and treat
Time from diagnosis to art start	5	NA	0.233	D4 Pt /Clinic Interface	4	this variable has changed with new guideline test and treat
Weeks clients considered lost to follow up are traced	0	NA		D4 Pt /Clinic Interface	4	missing for one site
Service Satisfaction - Care is not good	0	0	(no event)	D4 Pt /Clinic Interface	4	not enough cases
Service Satisfaction - Staff is disrespectfull	0	0	(no event)	D4 Pt /Clinic Interface	4	not enough cases
Mode of transport to clinic	0	NA	0.275	D4 Transport	1	documented impact on adherence/clinic visit
Has treatment supporter	0	472	0.146	D4 Treatment support	2	very low variability, most clients have treatment buddy
Attends support groups	0	82	0.263	D4 Treatment support	3	no association to outcome in study sample but in literature, and low cases
Clients per clinic day per clinical staff member	0	NA	0.000	D5 CTC Human Resources	1	This term was to describe the time availabel for direct patient management
Client per HCW	0	NA	0.000	D5 CTC Human Resources	2	this term includes non-medical staff and does not reflect clinic days
Clients per clinical staff member	0	NA	0.000	D5 CTC Human Resources	2	clinical knowledge might not be most important for sustainable service
All staff positions allocated to clinic	0	NA	0.000	D5 CTC Human Resources	2	this is a very braod term
All clinical staff positions allocated to clinic	0	NA	0.000	D5 CTC Human Resources	2	might be less informative than other HR variables
Costs to access clinic	0	NA	0.763	D5 Individual Resources	2	no association with outcome, positive bias can be expected as clients for whom this is relevant are likely not included in the study sample
Distance from household to clinic	0	NA	0.589	D5 Individual Resources	2	no association with outcome, positive bias can be expected as clients for whom this is relevant are likely not included in the study sample
Reason to miss clinic visit - lack of time	0	14	0.382	D5 Individual Resources	2	not enough cases

Reason to miss clinic visit - too far	0	10	0.664	D5 Individual Resources	2	not enough cases
Time to access clinic	1	NA	0.424	D5 Individual Resources	2	no association with outcome, positive bias can be expected as clients for whom this is relevant are likely not included in the study sample
Access to electricity in the HH	0	193	0.712	D5 Individual Resources	3	other variables have more evidence
Education	0	NA	0.412	D5 Individual Resources	3	little variance
Literacy	0	NA	0.529	D5 Individual Resources	3	other variables have more evidence
Profession	0	NA	0.668	D5 Individual Resources	3	large variance in the study sample
Reason to miss clinic visit - does not have enough money	0	0	(no event)	D5 Individual Resources	4	not enough cases
Ratio clients ever on ART to currently on ART	0	NA	0.000	D6 CTC Characteristics	1	rough parameter for attrition and positive bias
Years of operation	0	NA	0.000	D6 CTC Characteristics	1	shown to be relevant for treatment outcome in literature
Number of clients recieved as referral from lower health care levels	0	NA	0.000	D6 CTC Characteristics	1	would mean more sick patients
Number of clients recieved as up-referral from lower health care levels	0	NA	0.000	D6 CTC Characteristics	1	would mean less sick patients
Clients on ART at CTC	0	NA	0.000	D6 CTC Characteristics	1	easyer parameter than 1st/2nd line
Number of clients ever on ART at the clinic	0	NA	0.000	D6 CTC Characteristics	1	experience but also workload variable
Clinic days per patient managed	0	NA	0.000	D6 CTC Characteristics	2	other variables more specific
Clients on First Line	0	NA	0.000	D6 CTC Characteristics	2	better combined siteart
Number of clinic days per week	0	NA	0.036	D6 CTC Characteristics	2	Relevance depends on the number of clients seen in those days
Pt on Second Line	0	NA	0.000	D6 CTC Characteristics	4	could be a variable of service and difficult patient, but very low rate
Renovations conducted at the CTC	0	NA	0.029	D6 CTC Characteristics	4	just not a very meaningfull predictor for a model

Minor Renovations	0	NA	0.000	D6 CTC Characteristics	4	just not a very meaningfull predictor for a model
Major Renovations	0	NA	0.038	D6 CTC Characteristics	4	just not a very meaningfull predictor for a model
Marial status	0	NA	0.119	D6 Private Life	1	might be indication of overal resources available to participant
Number of financial dependants	0	NA	0.046	D6 Private Life	2	less variance due to categorical variables
Knows HIV status of spouse	0	309	0.883	D6 Private Life	3	not usefull to predict risk in single individuals
Number of adults in HH	0	NA	0.389	D6 Private Life	3	other variables have more evidence
Number of children in HH	0	NA	0.258	D6 Private Life	3	other variables have more evidence
Spouse is HIV positive	0	232	0.311	D6 Private Life	3	not usefull to predict risk in single individuals
Spouse is on ART	23	164	0.424	D6 Private Life	3	not usefull to predict risk in single individuals
Number of weeks of ARV the pharmacy typically keeps on hand	0	NA	0.000	D7 Pharmacy characteristics	4	not much variability
Pharmacy has access to electricity	0	NA	0.091	D7 Pharmacy characteristics	4	only MZRH has integrated CTC in OPD
Do non-pharmacy hospital personnel have access to medications stored in the pharmacy after hours (for example, nights/weekends)	0	NA	0.797	D7 Pharmacy characteristics	4	only MZRH has integrated CTC in OPD
Health Care Level	0	NA	0.000	D8 Hospital characteristics	1	Stratum, easy to assess
CTC integrated in OPD	0	NA	0.000	D8 Hospital characteristics	2	only MZRH has integrated CTC in OPD
Number of integrations of CTC with other disease specific clinics	0	NA	0.724	D8 Hospital characteristics	2	not much variance
Facility Size	0	NA	0.000	D8 Hospital characteristics	2	very borad
CTC integrated in TB clinic	0	NA	0.002	D8 Hospital characteristics	3	rather a facility characteristic
Location of Clinic	0	NA	0.420	D8 Hospital characteristics	3	not much variance
CTC integrated in ANC clinic	0	NA	0.272	D8 Hospital characteristics	4	only MZRH has integrated CTC in OPD
CD4 count at SV	0	NA	0.000	I0 Immunological control	1	highly associated with outcome

Lymphocyte count at SV	1	NA	0.003	I0 Immunological control	2	associated with outcome but less so than CD4 count and more missing values
Lymphocyte percentage at SV	1	NA	0.844	I0 Immunological control	2	associated with outcome but less so than CD4 count and more missing values
White blood count at SV	1	NA	0.000	I0 Immunological control	2	associated with outcome but less so than CD4 count and more missing values
Age	0	NA	0.083	I1 Biological determinants	1	known relevance, easy to assess
Gender	0	NA	0.877	I1 Biological determinants	1	known relevance, easy to assess
Blind at SV	0	16	0.757	I1 Biological determinants	4	not enough cases
Death at SV	0	5	0.642	I1 Biological determinants	4	not enough cases
Pregnant at SV	0	3	0.530	I1 Biological determinants	4	not enough cases
Years HIV infected	5	NA	0.149	I1 Biological determinants	4	mostly alligns with tonart, not easily to be verified when longer ART/confidential tests
ESR at SV	1	NA	0.005	I1 Immunologic reserve	1	could be a variable to well describe immune reconstitution
Plateltes at SV	1	NA	0.000	I1 Immunologic reserve	1	might be surrogate parameter of immunologic reserve
Hemoglobine % at SV	1	NA	0.141	I1 Immunologic reserve	2	Cd4 count considered more predictive
Hemoglobine at SV	1	NA	0.336	I1 Immunologic reserve	2	Cd4 count considered more predictive
Red blood cells at SV	1	NA	0.414	I1 Immunologic reserve	2	Cd4 count considered more predictive
Moncytes at SV	2	NA	0.652	I1 Immunologic reserve	2	Cd4 count considered more predictive
Monocytes percentage at SV	2	NA	0.042	I1 Immunologic reserve	2	Cd4 count considered more predictive
Neutrophiles % at SV	2	NA	0.777	I1 Immunologic reserve	2	Cd4 count considered more predictive
Neutrophiles at SV	2	NA	0.001	I1 Immunologic reserve	2	Cd4 count considered more predictive
Absoluet CD4 gain since ART start on SV	8	NA	0.000	I1 Immunologic reserve	2	BL variables expected to have little impact on current VL
CD4 count at BL	8	NA	0.160	I1 Immunologic reserve	2	BL variables expected to have little impact on current VL
Peak CD4 count	9	NA	0.108	I1 Immunologic reserve	2	BL variables expected to have little impact on current VL
Hemoglobine at BL	31	NA	0.617	I1 Immunologic reserve	2	BL variables expected to have little impact on current VL

Lymphocytes at BL	43	NA	0.133	I1 Immunologic reserve	2	BL variables expected to have little impact on current VL
Basophiles % at SV	42	NA	0.164	I1 Immunologic reserve	4	missing at sites
Basophiles at SV	42	NA	0.198	I1 Immunologic reserve	4	missing at sites
Eosinophiles at SV	42	NA	0.522	I1 Immunologic reserve	4	missing at sites
Eosinophiles % at SV	43	NA	0.214	I1 Immunologic reserve	4	missing at sites
WHO T-staging at SV	0	NA	0.687	I2 Co-Morbidity-Infectious	1	easy to assess at SV
Respiration rate	0	NA	0.309	I2 Co-Morbidity-Infectious	1	easy to assess at SV
Fever within the last three months or on SV	0	417	0.274	I2 Co-Morbidity-Infectious	2	not easy to collect in a standardized manner
Weight at SV	0	NA	0.267	I2 Co-Morbidity-Infectious	2	weight in itself might be good for AIDS but not for VL replication >400 copies/ml
BMI on SV	0	NA	0.123	I2 Co-Morbidity-Infectious	2	easy to assess at SV but only indirect connection to viraemia
Ever TB	37	113	0.383	I2 Co-Morbidity-Infectious	2	known relevane but few cases
Pain at SV	38	97	0.255	I2 Co-Morbidity-Infectious	2	to vague for standardized determination
Loss of Apetite within 3 months prior SV	38	52	0.158	I2 Co-Morbidity-Infectious	2	not easy to collect in a standardized manner
Night Sweats at SV	38	72	0.835	I2 Co-Morbidity-Infectious	2	other variables better indication of infection
Sore Throat at SV	38	53	0.275	I2 Co-Morbidity-Infectious	2	other variables better indication of infection
WHO3-Anaemia at SV	0	702	0.642	I2 Co-Morbidity-Infectious	3	single disease entity to specific for model
WHO2- Minor Mucocutaneous Manifestations at SV	72	196	0.088	I2 Co-Morbidity-Infectious	3	single disease entity to specific for model
WHO2 - Weight Loss < 10% at SV	75	178	0.472	I2 Co-Morbidity-Infectious	3	single disease entity to specific for model
WHO3-Weight loss >10% of Body Weight at SV	75	177	0.463	I2 Co-Morbidity-Infectious	3	single disease entity to specific for model
WHO3- Unexplained Chronic Diarrhea at SV	78	152	0.413	I2 Co-Morbidity-Infectious	3	single disease entity to specific for model
WHO2 - Recurrent Upper Respiratory Track Infections at SV	79	144	0.608	I2 Co-Morbidity-Infectious	3	single disease entity to specific for model

WHO2 - Herpes Zoster (within last 5 years) at SV	88	85	0.331	I2 Co-Morbidity-Infectious	3	single disease entity to specific for model
WHO3-TB at SV	88	83	0.352	I2 Co-Morbidity-Infectious	3	single disease entity to specific for model
WHO3-Severe Bacterial Infections at SV	89	76	0.537	I2 Co-Morbidity-Infectious	3	single disease entity to specific for model
WHO3-Candidiasis oral at SV	92	59	0.279	I2 Co-Morbidity-Infectious	3	single disease entity to specific for model
Temperature at SV	0	NA	0.486	I2 Co-Morbidity-Infectious	4	not enough cases with elevated temperature at SV
WHO3-Unexplained Prolonged Fever at SV	53	2	0.009	I2 Co-Morbidity-Infectious	4	not enough cases
WHO4-HIV Wasting Syndrome at SV	96	28	0.186	I2 Co-Morbidity-Infectious	4	not enough cases
WHO4-Disseminated Candidiasis at SV	97	18	(no event)	I2 Co-Morbidity-Infectious	4	not enough cases
WHO4-Disseminated TB at SV	98	14	0.880	I2 Co-Morbidity-Infectious	4	not enough cases
WHO4-Kaposi’s Sarcoma at SV	98	11	0.343	I2 Co-Morbidity-Infectious	4	not enough cases
WHO4-PneumocysticCarinii Pneumonia (PCP) at SV	98	11	(no event)	I2 Co-Morbidity-Infectious	4	not enough cases
WHO4-Cryptococcal, extrapulmonary at SV	99	8	(no event)	I2 Co-Morbidity-Infectious	4	not enough cases
WHO4-Encephalopathy at SV	99	5	(no event)	I2 Co-Morbidity-Infectious	4	not enough cases
WHO3-Oral Hairy Leukoplakia at SV	99	4	(no event)	I2 Co-Morbidity-Infectious	4	not enough cases
WHO4-Cryptosporidiosis with Diarrhea at SV	100	1	(no event)	I2 Co-Morbidity-Infectious	4	not enough cases

WHO4-Herpes simplex chronic at SV	100	1	(no event)	I2 Co-Morbidity-Infectious	4	not enough cases
Diastolic Blood Pressure at SV	0	NA	0.078	I2 Morbidity other	2	easy to assess at SV
Pulse at SV	0	NA	0.160	I2 Morbidity other	2	might indicate underlying infection but temperature would be better
Systolic Blood Pressure at SV	0	NA	0.019	I2 Morbidity other	2	easy to assess at SV
Asthma at SV	0	4	0.018	I2 Morbidity other	4	not enough cases
Ever had Asthma	0	4	0.069	I2 Morbidity other	4	not enough cases
Ever had Heart Disease	37	8	0.296	I2 Morbidity other	4	not enough cases
Heart Disease at SV	37	5	0.420	I2 Morbidity other	4	not enough cases
BMI at Baseline	3	NA	0.436	I3 Medical history	2	summary measure prefered over individual disease for model
WHO3-History of Unexplained Prolonged Fever at SV	53	331	0.699	I3 Medical history	2	summary measure prefered over individual disease for model
WHO2- History of Minor Mucocutaneous Manifestations	72	196	0.516	I3 Medical history	2	summary measure prefered over individual disease for model
WHO2-History of Weight Loss < 10% of Body Weight	75	177	0.736	I3 Medical history	2	summary measure prefered over individual disease for model
WHO3-Histroy of Weight loss >10% of Body Weight	75	177	0.368	I3 Medical history	2	summary measure prefered over individual disease for model
WHO3- History of Unexplained Chronic Diarrhea	78	152	0.686	I3 Medical history	2	summary measure prefered over individual disease for model
WHO2 - History of Recurrent Upper Respiratory Track Infections	79	144	0.051	I3 Medical history	2	summary measure prefered over individual disease for model
WHO3-History of TB	88	84	0.352	I3 Medical history	2	summary measure prefered over individual disease for model

WHO3- History of Severe Bacterial Infections	89	76	0.537	I3 Medical history	2	summary measure prefered over individual disease for model
WHO3- History of Anaemia at SV	94	43	0.536	I3 Medical history	2	summary measure prefered over individual disease for model
WHO4-History of HIV Wasting Syndrome at SV	96	28	0.237	I3 Medical history	2	summary measure prefered over individual disease for model
Clinical fialure at SV	0	21	0.981	I3 Medical history	3	low number of cases and no association with outcome, known bad predictive value
WHO Stage at BL	0	NA	0.328	I3 Medical history	3	Model should not include variables that can not retrospectively be assessed if missing on SV
Weight at BL	3	NA	0.337	I3 Medical history	3	Model should not include variables that can not retrospectively be assessed if missing on SV
Ever recieved PMTCT	0	30	0.578	I3 Medical history	4	not enough cases
Immune Reconstitution Syndrome	0	8	0.260	I3 Medical history	4	not enough cases
WHO Stage 1 - Persistent Generalized Lymphadenopathy	0	8	0.828	I3 Medical history	4	not enough cases
Ever had cancer	37	4	0.420	I3 Medical history	4	not enough cases
Ever had diabetis	37	8	0.700	I3 Medical history	4	not enough cases
Creatinine at BL	73	NA	0.233	I3 Medical history	4	very limited data
WHO2 - History of Herpes Zoster (within last 5 years)	88	85	0.331	I3 Medical history	4	not enough cases
WHO Stage 3 during or before ART - Candidiasis–Oral (Thrush)	92	59	0.065	I3 Medical history	4	not enough cases
WHO4-History of Disseminated Candidiasis	97	18	0.099	I3 Medical history	4	not enough cases
WHO4-History of Disseminated TB	98	15	0.898	I3 Medical history	4	not enough cases
WHO4-History of Kaposi’s Sarcoma at SV	98	11	0.047	I3 Medical history	4	not enough cases
WHO4-History of PneumocysticCarinii Pneumonia (PCP)	98	11	0.571	I3 Medical history	4	not enough cases

WHO4-History of Cryptococcal, extrapulmonary	99	8	0.635	I3 Medical history	4	not enough cases
WHO4- History of Encephalopathy	99	5	0.428	I3 Medical history	4	not enough cases
WHO3- History of Oral Hairy Leukoplakia	99	4	0.642	I3 Medical history	4	not enough cases
WHO4-History of Toxoplasmosis, CNS	100	3	0.386	I3 Medical history	4	not enough cases
WHO4- History of Herpes simplex chronic	100	1	0.657	I3 Medical history	4	not enough cases
WHO4-History of Cryptosporidiosis with Diarrhea	100	1	0.657	I3 Medical history	4	not enough cases

13.5 Case Report Forms

13.5.1 Site Observation Questionnaire

13.5.2 Nurse Administered Questionnaire

13.5.3 Participant Specific Case Report Form