University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2017 *Paper* 357

Evaluation of Progress Towards the UNAIDS 90-90-90 HIV Care Cascade: A Description of Statistical Methods Used in an Interim Analysis of the Intervention Communities in the SEARCH Study

Laura Balzer* Joshua Schwab†
Mark J. van der Laan‡ Maya L. Petersen**

*Department of Biostatistics, Harvard T.H. Chan School of Public Heath, lbbalzer@hsph.harvard.edu

†Division of Biostatistics, University of California, Berkeley, joshuaschwab@yahoo.com

‡Division of Biostatistics, University of California, Berkeley, laan@berkeley.edu

**Division of Biostatistics, University of California, Berkeley, mayaliv@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper357

Evaluation of Progress Towards the UNAIDS 90-90-90 HIV Care Cascade: A Description of Statistical Methods Used in an Interim Analysis of the Intervention Communities in the SEARCH Study

Laura Balzer, Joshua Schwab, Mark J. van der Laan, and Maya L. Petersen

Abstract

WHO guidelines call for universal antiretroviral treatment, and UNAIDS has set a global target to virally suppress most HIV-positive individuals. Accurate estimates of population-level coverage at each step of the HIV care cascade (testing, treatment, and viral suppression) are needed to assess the effectiveness of "test and treat" strategies implemented to achieve this goal. The data available to inform such estimates, however, are susceptible to informative missingness: the number of HIV-positive individuals in a population is unknown; individuals tested for HIV may not be representative of those whom a testing intervention fails to reach, and HIV-positive individuals with a viral load measured may not be representative of those for whom no viral load is obtained. We provide an in-depth description of the statistical methods (target parameters, assumptions, statistical estimands, and algorithms) used in an interim analysis of the intervention arm of the SEARCH Study (NCT01864603) to analyze progress towards the UNAIDS 90-90-90 target at study baseline and after one and two years. We describe the methods used to account for informative measurement in all analyses as well as for informative censoring in longitudinal analyses. We use targeted maximum likelihood estimation (TMLE) with Super Learning to generate semi-parametric efficient and double robust estimates of the care cascade among a open cohort of prevalent HIV-positive adults and among a closed cohort of baseline HIV-positive adults. TMLE is also used to evaluate predictors of poor outcomes.

Contents

1 Overview

In 2014, UNAIDS issued an ambitious new target for the HIV "care cascade" worldwide: by 2020, at least 90% of HIV-positive individuals should be diagnosed, at least 90% of those diagnosed should be receiving antiretroviral therapy (ART), and at least 90% of those on ART should have suppressed viral replication, for an overall target of 73% of all HIV-positive individuals virally suppressed (UNAIDS, 2014). The Sustainable East Africa Research in Community Health (SEARCH) Study is a cluster randomized trial to evaluate the health, economic, and educational impacts of a "Universal Test and Treat" HIV intervention compared to the country standard-of-care in 32 pair-matched communities in rural Kenya and Uganda (NCT01864603). At study baseline, a household census was used to enumerate all community residents. Immediately following the baseline census and annually thereafter, population-wide HIV testing was conducted in the intervention communities through a hybrid testing model consisting of a multi-disease community health campaign (CHC) followed by tracking and home-based testing for all enumerated residents who did not attend the CHC (Chamie et al., 2016). Plasma HIV RNA levels (viral loads) were also measured on all HIV-positive adults during these hybrid testing campaigns. We estimated progress towards the UNAIDS care cascade target among adult ($\geq$ 15 years of age) residents of the SEARCH intervention communities at study baseline and after one and two years of the intervention. We also evaluated the extent to which specific subgroups were at risk of poor cascade outcomes. In this document, we provide an in-depth description of the statistical methods (target parameters, assumptions, statistical estimands, and algorithms) used to generate these estimates. Full `R` code is available at `https://github.com/LauraBalzer/Estimating-90-90-90-in-SEARCH`.

We conducted three sets of complimentary analyses. First, as described in Section 3, we evaluated an open cohort of prevalent HIV-positive adult community residents at the time of the three annual rounds of population-wide HIV antibody and plasma HIV RNA level testing. In this open cohort, we estimated: 1) the proportion of HIV-positive individuals who had been previously diagnosed with HIV; 2) the proportion of previously diagnosed HIV-positive individuals who had previously been or were currently treated with ART; 3) the proportion of previously or currently treated HIV-positive individuals who were virally suppressed (viral load $<$ 500 copies/ml); and 4) the proportion of all HIV-positive individuals who were virally suppressed. In these analyses, the number of HIV-positive individuals was estimated accounting for incomplete and possibly informative HIV testing coverage, and the number of HIV-positive individuals with viral suppression was estimated accounting for incomplete and possibly informative viral load testing coverage. We repeated these analyses within *a priori*-specified subgroups of sex, age, and country. Section 3.6 provides a worked example demonstrating the approach used to adjust for missing measures.

Second, as described in Section 4, we evaluated a closed cohort of adult residents with an HIV diagnosis at or before study baseline. In this cohort, we estimated the proportion who at baseline and after one and two years 1) had died; 2) had out-migrated from the community; 3) were newly diagnosed; 4) had never initiated ART; 5) had initiated but were unsuppressed; and 6) had initiated ART and were virally suppressed. In these analyses, we adjusted for incomplete and possibly informative viral load testing coverage among HIV-positive residents. We repeated these analyses within *a priori*-specified subgroups of sex, age, and country.

Third, as described in Section 5, we again analyzed the closed cohort of adult residents with an HIV diagnosis at or before study baseline. We estimated the proportion of this cohort who were virally unsuppressed after one and two years, while treating death and out-migration as right-censoring events. We adjusted for potentially informative censoring in addition to potentially informative missing viral load measures. We repeated these analyses within *a priori*-specified subgroups defined by baseline prior HIV diagnosis, ART use, and viral suppression. In the same cohort of baseline HIV-positive adults, analogous analyses were performed to estimate the proportion of the cohort who had never initiated ART. In an expanded closed cohort of all adult community residents without an HIV diagnosis prior to baseline test-

ing, analogous analyses were performed to estimate the proportion of residents who were tested for HIV at least once during follow-up. For each of these outcomes, we further evaluated univariable and multivariable associations between demographic factors and poor cascade outcomes (failing to suppress viral replication, never initiating ART, and never testing for HIV, respectively).

2 Observed data

We consider time points $t = \{0, 1, 2\}$, corresponding to the community-specific dates of the annual population-wide HIV and viral load testing rounds at study baseline, follow-up year 1, and follow-up year 2, respectively.

- Let B denote baseline variables measured on all individuals during the household census enumeration conducted immediately prior to the first round of testing. These include age, sex, marital status, education, occupation, mobility, household wealth index, community of residence, and country.
- Let D_t be an indicator that an individual has died by time t, and M_t be an indicator that an individual has migrated out of the community by time t. We define C_t as an indicator of censoring by death or out-migration by time t. At baseline, $D_0 = M_0 = C_0 = 0$ deterministically.
- Let Y_t^* and VLV_t^* denote the underlying values of HIV serostatus and viral load at time t, irrespective of whether these values are measured. Further, let $Supp_t^*$ be an indicator that an individual's viral load at time t is <500 copies/ml: $Supp_t^* = \mathbb{I}(VLV_t^* < 500)$. Throughout, we use viral loads measured at CHC/tracking rather than those measured at clinic to minimize the potential for informative missingess, as in-clinic measures inherently depend on an individual's retention in HIV care.
- Let pDx_t be an observed indicator that an HIV-positive individual has been diagnosed prior to time t. Let $eART_t$ be an observed indicator that an HIV-positive individual has ever initiated ART prior to time t. Both variables are step functions, jumping to one as soon as there is evidence of prior diagnosis or ART use, respectively, and remaining equal to one thereafter. Classification of an individual as previously diagnosed or previously treated requires either health record documentation or a suppressed viral load in an individual documented to be HIV-positive. If an individual does not have evidence of prior diagnosis, then he or she is assumed not be previously diagnosed as HIV-positive. Likewise, if an individual does not have evidence of ART use, then he or she is assumed not to have initiated ART.
- Let CHC_t and Tr_t denote indicators that an individual was seen at a CHC or at subsequent tracking at time t, respectively.
- Let $TstHIV_t$ denote an indicator that HIV status is known at time t. HIV status is considered "known" if an individual was tested for HIV by SEARCH at time t or already had a known HIV-positive status from previous health records or documented tests.
- Let Δ_t denote an indicator that a individual was contacted at the CHC/tracking and had a known HIV status at time t: $\Delta_t = \mathbb{I}[(CHC_t = 1 \text{ or } Tr_t = 1)\ \&\ TstHIV_t = 1]$.
- Let $Y_t = TstHIV_t \times Y_t^*$ denote observed HIV serostatus at time t.
- Let $TstVL_t$ denote an indicator that a viral load was measured at CHC/tracking at time t.
- Let $Supp_t = TstVL_t \times Supp_t^*$ denote observed viral suppression (as measured at CHC/tracking) at time t.

- Let $ever.supp_t$ be an indicator of having at least one measured viral load < 500 copies/ml at or after study baseline. This variable uses additional viral load measures made in the clinic, beyond the annual measures made at CHC/tracking. This variable treats the absence of a measured suppressed viral load as evidence of failure to suppress.

The observed data on a given individual at time point t are

$$\begin{aligned} O_t =&(D_t, M_t, pDx_t, eART_t, CHC_t, Tr_t, TstHIV_t, \Delta_t, \\ &TstHIV_t \times Y_t^*, TstVL_t, TstVL_t \times Supp_t^*, ever.supp_t) \\ =&(C_t, pDx_t, eART_t, CHC_t, Tr_t, TstHIV_t, \Delta_t, \\ &Y_t, TstVL_t, Supp_t, ever.supp_t). \end{aligned}$$

All variables in O_t are indicator variables. The observed data on a given individual consist of

$$O = (B, O_0, O_1, O_2)$$

where for ease of notation, we assume that variables after censoring by death or out-migration are equal to their last observed values. We use overbars to denote a variable's history (e.g. $\bar{O}_t \equiv (O_0, ..., O_t)$ for $t > 0$).

Throughout we make use of the following relationships between variables. One can only have a previous diagnosis if HIV-positive: $\mathbb{P}(pDx_t = 1|Y_t^* = 0) = 0$. One can only initiate ART after receiving a diagnosis: $\mathbb{P}(eART_t = 1|pDx_t = 0) = 0$. We also assume no HIV-positive individuals control viral replication below 500 copies/ml in the absence of treatment: $\mathbb{P}(Supp_t^* = 1|eART_t = 0, Y_t^* = 1) = 0$.

3 Analysis of the open cohort of prevalent HIV-positive adults

3.1 Overview

For three time points $t = \{0, 1, 2\}$, we aim to estimate cascade coverage and population-level viral suppression in an open cohort consisting of all prevalent HIV-positive adult residents of the community at that time point. In other words, our t-specific target population is all individuals who are alive, not out-migrated, ≥ 15 years of age at time t, and HIV-positive at time t. All parameters below are thus conditional on $D_t = M_t = 0$, and $age_t \geq 15$ (in addition to $Y_t^* = 1$). We suppress the former conditioning in our notation to simplify presentation. In the primary analysis, we restrict our target population to baseline enumerated stable (≥ 6 months of past year in the community during the census) residents. In sensitivity analyses, we include non-stable residents and in-migrants.

When estimating serial cascade coverage in this open cohort, we leverage the full longitudinal data structure to adjust for potentially informative missing measures of HIV status and viral load. For example, individuals testing for HIV at baseline may also be more likely to be successfully contacted at CHC or tracking in subsequent years than individuals who were not tested at baseline, potentially inflating cascade coverage estimates unless we adjust for past attendance.

As detailed in the following sections, our general approach is to identify and estimate the following four population-level proportions:

1. **HIV prevalence**:
$$\mathbb{P}(Y_t^* = 1)$$
2. **Proportion who are HIV-positive and previously diagnosed**:
$$\mathbb{P}(pDx_t = 1, Y_t^* = 1)$$
3. **Proportion who are HIV-positive, previously diagnosed, and ever on ART**:
$$\mathbb{P}(eART_t = 1, pDx_t = 1, Y_t^* = 1)$$

4. **Proportion who are HIV-positive, previously diagnosed, ever on ART, and virally suppressed**:

$$\mathbb{P}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, Y_t^* = 1)$$

As explicitly discussed below, probabilities #2 and #3 are estimated as empirical proportions, while probabilities #1 and #4 are estimated with targeted maximum likelihood estimation (TMLE), a double robust and semi-parametric efficient approach that allows adjustment for missing measurements and censoring (van der Laan and Rubin, 2006; van der Laan and Rose, 2011). To minimize model misspecification bias and optimize estimator performance, nuisance parameters are estimated using Super Learner, an ensemble machine learning method (van der Laan et al., 2007). When implementing Super Learner, we use 5-fold cross-validation and specify an algorithm library consisting of general additive models, stepwise regression, and logistic regression, all with and without pre-screening based on univariate outcome correlations. Estimators are implemented using the `ltmle` v0.9-8-4 (Schwab et al., 2016) and `SuperLearner` v2.0-21 packages (Polley and van der Laan, 2014) in `R` (R Core Team, 2015). Code to implement these estimators is available at `https://github.com/LauraBalzer/Estimating-90-90-90-in-SEARCH`.

Given estimates of probabilities #1-4, our population-level estimates of the 90-90-90 cascade and viral suppression are obtained by taking the following ratios.

- **Proportion of all HIV-positive individuals who have a prior diagnosis at time t**:

$$\mathbb{P}(pDx_t = 1 \mid Y_t^* = 1) = \frac{\mathbb{P}(pDx_t = 1, Y_t^* = 1)}{\mathbb{P}(Y_t^* = 1)}.$$

- **Proportion of all HIV-positive individuals with a prior diagnosis who have ever started ART at time t:**

$$\mathbb{P}(eART_t = 1 \mid pDx_t = 1, Y_t^* = 1) = \frac{\mathbb{P}(eART_t = 1, pDx_t = 1, Y_t^* = 1)}{\mathbb{P}(pDx_t = 1, Y_t^* = 1)}.$$

- **Proportion of all HIV-positive individuals with prior ART initiation who are virally suppressed at time t**:

$$\mathbb{P}(Supp_t^* = 1 \mid eART_t = 1, pDx_t = 1, Y_t^* = 1) \quad = \frac{\mathbb{P}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, Y_t^* = 1)}{\mathbb{P}(eART_t = 1, pDx_t = 1, Y_t^* = 1)}.$$

- **Proportion of all HIV-positive individuals who are virally suppressed at time t:**

$$\mathbb{P}(Supp_t^* = 1 \mid Y_t^* = 1) = \frac{\mathbb{P}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, Y_t^* = 1)}{\mathbb{P}(Y_t^* = 1)}.$$

In unadjusted (secondary) analyses, we also implement simple estimators of each cascade step, equivalent to the empirical proportions among individuals who have measured values and are contacted at the CHC/tracking (Section 3.3). We also refer the reader to Section 3.6 for a worked example of our approach and illustration of why other methods (e.g. unadjusted estimators) might fall short. Statistical inference, including Wald-Type 95% confidence intervals, are based on influence curve standard error estimators that treat households as the unit of independence (van der Laan and Rubin, 2006; van der Laan and Rose, 2011).

3.2 Primary analysis

3.2.1 HIV prevalence

Our goal is to identify and then estimate the proportion of the target population who are HIV-positive at time t:

$$\mathbb{P}(Y_t^* = 1).$$

To identify this parameter, we assume that within strata defined by baseline covariates, past testing, and cascade history, the prevalence of HIV among those seen at CHC/tracking is representative of the prevalence of HIV among those not seen:

$$\Delta_t \perp\!\!\!\perp Y_t^* \mid B, \bar{O}_{t-1}.$$

As our missingess indicator, we use Δ_t rather than $TstHIV_t$, because $TstHIV_t$ is a direct function of underlying HIV status Y_t^*. A previously diagnosed HIV-positive individual's status can be ascertained through records without attending the CHC/tracking, whereas an HIV-negative individual's status cannot be.

By definition, the above assumption holds for individuals known to be HIV-positive at the previous time point (i.e. with $Y_{t-1} = 1$):

$$\mathbb{P}(Y_t^* = 1 | Y_{t-1} = 1) \ = \ 1.$$

Furthermore, among those not previously known to be HIV-positive ($Y_{t-1} = 0$), there is only variability in baseline demographics, past CHC/tracking attendance, and past HIV testing history. Therefore, for the subgroup of individuals without a prior HIV diagnosis by the close of the previous year's testing ($Y_{t-1} = 0$), we assume that within strata defined by baseline demographics, past CHC/tracking contact, and HIV testing history (e.g. number of prior negative tests), HIV prevalence among those tested at CHC/tracking is representative of HIV prevalence among those not tested:

$$\Delta_t \perp\!\!\!\perp Y_t^* \mid Y_{t-1} = 0, C\bar{H}C_{t-1}, \bar{T}r_{t-1}, Tst\bar{H}IV_{t-1}, B.$$

We further assume positivity: there are no strata (defined by baseline demographics, CHC/tracking history, and testing history) in which zero previously undiagnosed individuals are tested at the CHC/tracking at time t:

$$\mathbb{P}(\Delta_t = 1 | Y_{t-1} = 0, C\bar{H}C_{t-1}, \bar{T}r_{t-1}, Tst\bar{H}IV_{t-1}, B) > 0.$$

Let us denote these strata as $L_t \equiv (C\bar{H}C_{t-1}, \bar{T}r_{t-1}, Tst\bar{H}IV_{t-1}, B)$. Under the above assumptions and using the above deterministic knowledge, we have the following identifiability result:

$$\begin{aligned}
\mathbb{P}(Y_t^* = 1) &= \mathbb{P}(Y_t^* = 1 | Y_{t-1} = 1)\mathbb{P}(Y_{t-1} = 1) + \mathbb{P}(Y_t^* = 1 | Y_{t-1} = 0)\mathbb{P}(Y_{t-1} = 0) \\
&= \mathbb{P}(Y_{t-1} = 1) \\
&\quad + \mathbb{P}(Y_{t-1} = 0) \sum_{l_t} \big[\mathbb{P}(Y_t = 1 | \Delta_t = 1, L_t = l_t, Y_{t-1} = 0)\mathbb{P}(L_t = l_t | Y_{t-1} = 0)\big].
\end{aligned}$$

Our estimand can be interpreted as the proportion known to be HIV-positive at the prior time point plus the adjusted proportion not previously known to be HIV-positive who are known to be HIV-positive at t. The latter accounts for incomplete measurement of HIV status among this group. We stratify the population not previously known to be HIV-positive on baseline demographics, past CHC/tracking attendance, and HIV testing history; assume that within each stratum, HIV prevalence among individuals tested at the CHC/tracking is the representative of prevalence among those without an HIV test result, and then combine these stratum-specific estimates into a single standardized estimate.

For estimation, we use TMLE with Δ_t as the single intervention node, with (Y_{t-1}, L_t) as the adjustment set, and knowledge that $\mathbb{P}(Y_t = 1 | \Delta_t = 1, L_t = l_t, Y_{t-1} = 1) = 1$ for individuals known to be HIV-positive at the close of the prior round of testing. The estimated number of prevalent HIV-positive individuals is calculated as the estimated prevalence times the target population size at time t.

3.2.2 Proportion HIV-positive, and previously diagnosed

Our goal is to identify and then estimate the proportion of the target population who are previously diagnosed with HIV at time t.

$$\mathbb{P}(pDx_t = 1, Y_t^* = 1) = \mathbb{P}(pDx_t),$$

where we use that prior diagnosis of HIV implies that an individual is HIV-positive at time t. Estimation is based on the empirical proportion of the target population with evidence of a prior diagnosis at time t: a positive HIV test prior to time t, a record of HIV care prior to time t, or a viral load < 500 copies/ml at or before time t among individuals who are confirmed HIV-positive. The number of previously diagnosed HIV-positive individuals is calculated as the proportion with evidence of a prior diagnosis times the target population size at time t.

Failing to identify records of prior diagnosis among HIV-positive individuals will result in underestimation of prior diagnosis. This motivates our use of a suppressed viral load in a confirmed HIV-positive individual as evidence of being on ART at that time point and thus having been previously diagnosed. However, this approach assumes that no HIV-positive individuals control viral replication below 500 copies/ml in the absence of treatment, and results in potentially differential reduction in misclassification (missed ART records are only corrected among those individuals with suppressed viral loads). In secondary analyses, we employ an alternative assumption that prior diagnosis among HIV-positive individuals seen at the CHC/tracking is representative of prior diagnosis among HIV-positive individuals not seen. Finally, in recognition of the increased potential for missing records of prior diagnosis at study baseline, we conduct an additional sensitivity analysis in which self-report is used as evidence of prior diagnosis.

3.2.3 Proportion HIV-positive, previously diagnosed, and ever on ART

Our goal is to identify and then estimate the proportion of the target population who are HIV-positive, previously diagnosed, and have initiated ART by time t.

$$\mathbb{P}(eART_t = 1, pDx_t = 1, Y_t^* = 1) = \mathbb{P}(eART_t = 1)$$

where we have used that ART initiation implies that an individual has been previously diagnosed and is HIV-positive. Estimation is based on the empirical proportion of the target population who have evidence ART initiation by time t: a health care record with an ART start date prior to time t, or a viral load < 500 copies/ml at or before time t among individuals who are confirmed HIV-positive. The number of HIV-positive individuals ever on ART is calculated as the proportion with evidence of prior ART initiation times the target population size as time t.

Failing to identify records of ART use among HIV-positive individuals will result in underestimation of the proportion ever on ART. This motivates our use of a suppressed viral load in a confirmed HIV-positive individual as evidence of being on ART at that time point. Again, this approach assumes that no HIV-positive individuals control viral replication below 500 copies/ml in the absence of treatment, and results in potentially differential reduction in misclassification (missed ART records are only corrected among those individuals with suppressed viral loads). In secondary analyses, we employ an alternative assumption that ART use among HIV-positive individuals seen at the CHC/tracking is representative of use among HIV-positive individuals not seen.

3.2.4 Proportion HIV-positive, previously diagnosed, ever on ART, and virally suppressed

Our goal is to identify and then estimate the proportion of the target population who are HIV-positive, previously diagnosed, ever on ART, and virally suppressed at time t:

$$\mathbb{P}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, Y_t^* = 1).$$

Now we have missing measures on both HIV status and viral load. To simplify notation, let

$$Z_t^* \equiv \mathbb{I}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, Y_t^* = 1)$$

denote a binary indicator the outcome of interest, and let

$$Z_t \equiv \mathbb{I}(Supp_t = 1, eART_t = 1, pDx_t = 1, Y_t = 1)$$

be its observed analog.

We formulate the identifiability of our target parameter $\mathbb{P}(Z_t^* = 1)$ as a longitudinal dynamic regime (e.g. Hernán et al. (2006); van der Laan and Petersen (2007); Robins et al. (2008)). Other approaches are also possible, and several alternative formulations result in the same estimand. Here, we consider a two component hypothetical intervention on the missingness mechanism:

- $d0$: set $\Delta_t = 1$. In other words, ensure that all individuals attend the CHC or tracking at time t and have known HIV status.
- $d1(Y_t)$: if $(Y_t = 1)$, set $TstVL_t = 1$; else set $TstVL_t = 0$. In other words, if an individual is known to be HIV-positive at time t, ensure that his or her viral load is measured.

Under this hypothetical joint intervention, we would have complete measurement of underlying HIV status and complete measurement of viral loads among all HIV-positive individuals.

To identify the probability of suppression under such a hypothetical intervention, we assume the following (sequential) randomization assumptions (Robins, 1986):

$$\begin{aligned} Z_t^* &\perp\!\!\!\perp \Delta_t \mid L_t \\ Z_t^* &\perp\!\!\!\perp TstVL_t \mid Y_t, \Delta_t = 1, L_t \end{aligned} \tag{1}$$

where $L_t \equiv (B, \bar{O}_{t-1}, pDx_t, eART_t)$. Together with the corresponding positivity assumptions, our target parameter is identified as

$$\begin{aligned} \mathbb{P}(Z_t^* = 1) &= \sum_{l_t, y_t} \mathbb{P}(Z_t = 1 \big| TstVL_t = d1(y_t), Y_t = y_t, \Delta_t = 1, L_t = l_t) \mathbb{P}(Y_t = y_t \big| \Delta_t = 1, L_t = l_t) \mathbb{P}(L_t = l_t) \\ &= \sum_{b, \bar{o}_{t-1}} \mathbb{P}(Supp_t = 1 \big| TstVL_t = 1, Y_t = 1, \Delta_t = 1, eART_t = 1, pDx_t = 1, \bar{o}_{t-1}, b) \\ &\quad \times \mathbb{P}(Y_t = 1 \big| \Delta_t = 1, eART_t = 1, pDx_t = 1, \bar{o}_{t-1}, b) \\ &\quad \times \mathbb{P}(eART_t = 1, pDx_t = 1, \bar{o}_{t-1}, b). \end{aligned}$$

For the second equality, we use that the joint outcome Z_t is deterministically 0 if $Y_t = 0$ OR $pDx_t = 0$ OR $eART_t = 0$. Also in the second equality, we use our definition of dynamic treatment rule $d1(Y_t = 1) = 1$. The final equality is now in terms of the observed data distribution and can be estimated with longitudinal TMLE with a dynamic treatment regime.

Instead, we simplify our approach by noting the following. Both randomization assumptions (Eq. 1) hold deterministically among individuals never on ART ($eART_t = 0$). Likewise, ever ART use ($eART_t = 1$) implies prior diagnosis ($pDx_t = 1$) and an HIV-positive status ($Y_t^* = 1$). Finally, having a viral load measured at the CHC/tracking $TstVL_t = 1$ implies that the individual was known to be HIV-positive and attended the CHC/tracking $\Delta_t = 1$. Therefore, our target parameter is identified[1] as

$$\begin{aligned} \mathbb{P}(Z_t^* = 1) &= \sum_{b, \bar{o}_{t-1}} \mathbb{P}(Supp_t = 1 \big| TstVL_t = 1, eART_t = 1, \bar{o}_{t-1}, b) \times \mathbb{P}(eART_t = 1, \bar{o}_{t-1}, b) \\ &= \mathbb{P}(eART_t = 1) \times \sum_{b, \bar{o}_{t-1}} \mathbb{P}(Supp_t = 1 \big| TstVL_t = 1, eART_t = 1, \bar{o}_{t-1}, b) \\ &\quad \times \mathbb{P}(\bar{o}_{t-1}, b | eART_t = 1). \end{aligned}$$

[1] Our identifiability assumptions reduce to $Supp_t^* \perp\!\!\!\perp TstVL_t \big| eART_t = 1, \bar{O}_{t-1}, B$. For HIV-positive individuals who have initiated ART, we assume that within strata defined by baseline demographics, past HIV testing, and cascade history, suppression among those with a viral load measured at CHC/tracking is representative of suppression among those with a missing viral load. We assume the corresponding positivity assumption: $\mathbb{P}(TstVL_t = 1 \mid eART_t = 1, \bar{O}_{t-1}, B) > 0$. We require some positive probability of having viral load measured at the CHC/tracking, given the HIV-positive individual has initiated ART, regardless of baseline demographics or the observed past.

Thus, our estimand is the proportion of individuals known to have started ART multiplied by the adjusted probability of being suppressed given prior ART initiation. The latter is estimated by stratifying the population with prior ART initiation on baseline demographics and cascade history, assuming that within each stratum the proportion suppressed among those with a viral load test available is representative of the proportion suppressed without a viral load test, and combining these stratum-specific proportions into a single standardized estimate.

Estimation is based on a TMLE with a intervention node as $TstVL_t$ and adjustment set as $(eART_t, \bar{O}_{t-1}, B)$. During estimation, we use knowledge that the outcome is 0 if $eART_t = 0$. The estimated number of virally suppressed HIV-positive individuals is calculated as the estimated proportion suppressed times the target population size at time t.

3.2.5 The 90-90-90 cascade and population-level viral suppression

The previous subsections described our procedure for obtaining estimates of

1. The proportion of the target population who is HIV-positive: $\mathbb{P}(Y_t^* = 1)$
2. The proportion of the target population who is HIV-positive and previously diagnosed: $\mathbb{P}(pDx_t = 1, Y_t^* = 1)$
3. The proportion of the target population who is HIV-positive, previously diagnosed, and ever on ART: $\mathbb{P}(eART_t = 1, pDx_t = 1, Y_t^* = 1)$
4. The proportion of the target population who is HIV-positive, previously diagnosed, ever on ART, and virally suppressed: $\mathbb{P}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, Y_t^* = 1)$.

These estimates can be translated into estimates of the UNAIDS 90-90-90 cascade target by taking the following ratios. Inference is obtained by the Delta Method.

- **Proportion of all HIV-positive individuals who have a prior diagnosis at time t:**

$$\mathbb{P}(pDx_t = 1 \mid Y_t^* = 1) = \frac{\mathbb{P}(pDx_t = 1, Y_t^* = 1)}{\mathbb{P}(Y_t^* = 1)}.$$

- **Proportion of all HIV-positive individuals with a prior diagnosis who have ever started ART at time t:**

$$\mathbb{P}(eART_t = 1 \mid pDx_t = 1, Y_t^* = 1) = \frac{\mathbb{P}(eART_t = 1, pDx_t = 1, Y_t^* = 1)}{\mathbb{P}(pDx_t = 1, Y_t^* = 1)}.$$

- **Proportion of all HIV-positive individuals with prior ART initiation who are virally suppressed at time t:**

$$\mathbb{P}(Supp_t^* = 1 \mid eART_t = 1, pDx_t = 1, Y_t^* = 1) = \frac{\mathbb{P}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, Y_t^* = 1)}{\mathbb{P}(eART_t = 1, pDx_t = 1, Y_t^* = 1)}.$$

- **Proportion of all HIV-positive individuals who are virally suppressed at time t:**

$$\mathbb{P}(Supp_t^* = 1 \mid Y_t^* = 1) = \frac{\mathbb{P}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, Y_t^* = 1)}{\mathbb{P}(Y_t^* = 1)}.$$

3.3 Unadjusted secondary analysis

In an unadjusted analysis, we estimate each 90-90-90 cascade step and overall population-level suppression using simple empirical proportions among individuals seen at the CHC/tracking with known HIV status ($\Delta_t = 1$) and with viral load measured ($TstVL_t = 1$) for the suppression outcome. We make the following assumptions and note that several alternative formulations result in the same estimands.

1. We assume HIV prevalence among those with known HIV status seen at the CHC/tracking is representative of HIV prevalence in the target population:

$$Y_t^* \perp\!\!\!\perp \Delta_t.$$

2. In these secondary analyses, we assume that prior diagnosis is completely measured only among HIV-positive individuals seen at the CHC/tracking. We use pDx_t^* to denote underlying prior diagnoses and $pDx_t = \Delta_t \times pDx_t^*$ as its observed analog. We assume the proportion of individuals with a prior diagnosis seen at the CHC/tracking is representative of the proportion of individuals with a prior diagnosis in the target population:

$$pDx_t^* \perp\!\!\!\perp \Delta_t.$$

3. In these secondary analyses, we assume that prior ART use is completely measured only among HIV-positive individuals seen at the CHC/tracking. We use $eART_t^*$ to denote underlying prior ART use and $eART_t = \Delta_t \times eART_t^*$ as its observed analog. We assume the proportion of individuals seen at the CHC/tracking who have ever used ART is representative of the proportion of individuals who have ever used ART in the target population:

$$eART_t^* \perp\!\!\!\perp \Delta_t.$$

4. Finally, we assume that for HIV-positive individuals who have initiated ART, suppression among those with viral loads measured at the CHC/tracking is representative of suppression among those with missing viral load measures:

$$Supp_t^* \perp\!\!\!\perp TstVL_t \mid eART_t^* = 1.$$

Under these assumptions, we estimate cascade coverage and population-level suppression in the secondary analysis as follows.

1. **Proportion of all HIV-positive individuals who have a prior diagnosis at time t:** Under these assumptions and using that prior diagnosis implies an HIV-positive status, we have that

$$\mathbb{P}(pDx_t^* = 1 | Y_t^* = 1) = \frac{\mathbb{P}(pDx_t = 1 | \Delta_t = 1)}{\mathbb{P}(Y_t = 1 | \Delta_t = 1)}$$

The right hand side is estimated as the empirical proportion of individuals seen at CHC/tracking who have a prior diagnosis, divided by the empirical proportion of individuals seen at CHC/tracking who are known to be HIV positive. This is equivalent to the number of previously diagnosed HIV-positive individuals attending CHC/tracking, divided by the number of HIV-positive individuals attending CHC/tracking.

2. **Proportion of all HIV-positive individuals with a prior diagnosis who have ever started ART at time t**: Under these assumptions and using that prior ART initiation implies a prior diagnosis and an HIV-positive status, we have

$$\mathbb{P}(eART_t^* = 1 | pDx_t^* = 1, Y_t^* = 1) = \frac{\mathbb{P}(eART_t = 1 | \Delta_t = 1)}{\mathbb{P}(pDx_t = 1 | \Delta_t = 1)}.$$

The right hand side is estimated as the empirical proportion of individuals seen at CHC/tracking who have ever initiated ART, divided by the empirical proportion of individuals seen at CHC/tracking who have a prior diagnosis. This is equivalent to the number of HIV-positive individuals attending CHC/tracking who have ever initiated ART, divided by the number of HIV-positive individuals attending CHC/tracking who have a prior diagnosis.

3. **Proportion of all HIV-positive individuals with prior ART initiation who are virally suppressed at time** t: Under these assumptions and using that prior ART initiation implies a prior diagnosis and an HIV-positive status, we have

$$\begin{aligned}\mathbb{P}(Supp_t^* = 1 | eART_t^* = 1, pDx_t^* = 1, Y_t^* = 1) &= \mathbb{P}(Supp_t^* = 1 | TstVL_t = 1, eART_t^* = 1) \\ &= \mathbb{P}(Supp_t = 1 | TstVL_t = 1, eART_t = 1)\end{aligned}$$

(Recall having a viral load measured at the CHC/tracking $TstVL_t = 1$ implies the individual was known to be HIV-positive and attended the CHC/tracking $\Delta_t = 1$.) The right hand side is estimated as the empirical proportion of individuals with a measured viral load who are currently suppressed, divided by the empirical proportion of individuals with a measured viral load who have ever initiated ART. This is equivalent to the number of HIV-positive individuals with a measured viral load who are currently suppressed, divided by number of HIV-positive individuals with a measured viral load who ever initiated ART.

4. **Proportion of all HIV-positive individuals who are virally suppressed at time** t**:** As detailed in the worked example (Section 3.6), we can identify the proportion of all HIV-positive who are suppressed in this secondary analysis as

$$\mathbb{P}(Supp_t^* = 1 | Y_t^* = 1) = \mathbb{P}(Supp_t = 1 | TstVL_t = 1).$$

The right hand side is estimated with the empirical proportion of all HIV-positive individuals with a viral load measured at the CHC/tracking who are currently suppressed. This is equivalent to the number of HIV-positive individuals with observed viral suppression, divided by number of HIV-positive individuals with a measured viral load. See Section 3.6 for further details.

3.4 Stratified analysis

We estimate the same parameters as in Section 3.2, but now within the following strata defined by baseline variables included in B.

1. Sex (Male vs. Female)
2. Age (Younger: 15-24 years vs. Older: > 24 years) at time t
3. Country of residence (Uganda vs. Kenya)

3.5 Additional sensitivity analyses

We implement the following sensitivity analyses:

1. Expanding the target population to include non-stable residents and in-migrants
2. Including self-report as evidence of prior diagnosis at baseline

3.6 Worked example: estimating population-level viral suppression

To illustrate the approach taken in the primary analysis and detailed in Section 3.2, we provide the following worked example.

3.6.1 Target parameter

Our goal is to estimate the population-level viral suppression among HIV-positive adults:

$$\mathbb{P}(Supp_t^* = 1 \mid Y_t^* = 1) = \frac{\mathbb{P}(Supp_t^* = 1, Y_t^* = 1)}{\mathbb{P}(Y_t^* = 1)}.$$

The numerator is the joint probability of suppressed and being HIV-positive (irrespective of whether HIV status or viral load is measured), and the denominator is the underlying prevalence of HIV. We first focus on estimating the denominator and then the numerator. Finally, we combine these estimates to obtain an estimate of population-level viral suppression among HIV-positive adults.

3.6.2 Unadjusted (secondary) analysis approach

Denominator - HIV Prevalence: Our goal is to estimate population-level HIV prevalence: $\mathbb{P}(Y_t^* = 1)$. One approach would be to assume HIV prevalence is the same among those with known status and those with unknown status. Under this assumption, the estimated HIV prevalence at baseline ($t = 0$) is

$$\mathbb{P}(Y_0^* = 1) = \mathbb{P}(Y_0 = 1 | TstHIV_0 = 1) = \frac{\text{\# known to be HIV-positive at baseline}}{\text{population with known status at baseline}} = \frac{7108}{69283} = 10.3\%$$

This approach is potentially problematic for the following reasons. First, the status of an HIV-uninfected individual can only be ascertained through testing at the CHC/tracking, while the status of a previously diagnosed HIV-infected individual can be ascertained through health records without attending the CHC/tracking. This assumption is further problematic at later time points. For $t > 0$, we are more likely to know a previously HIV-infected individual's status due to both health records and because he/she only needed to be tested for HIV once. In other words, the missingness variable $TstHIV_t$ is a function of underlying HIV status Y_t^*. Therefore, in all analyses we rely on HIV status as ascertained at the CHC/tracking. This incorporates attendees who test for HIV (both with positive and negative results) and attendees who are previously diagnosed as HIV-infected (and thus not retested).

In the secondary analyses (Section 3.3), we assume that the prevalence of HIV among those testing at the CHC/tracking or attending the CHC/tracking with a documented prior HIV-positive test result ($\Delta_t = 1$) is representative of prevalence among those not ($\Delta_t = 0$). (We relax this assumption below.) With this assumption, we can identify HIV prevalence as

$$\mathbb{P}(Y_t^* = 1) \;=\; \mathbb{P}(Y_t = 1 | \Delta_t = 1) = \frac{\mathbb{P}(Y_t = 1, \Delta_t = 1)}{\mathbb{P}(\Delta_t = 1)}$$

By this approach (Table 1), the estimated baseline prevalence of HIV in SEARCH is

$$\frac{(\text{\# HIV-positive \& seen with known status})/77774}{(\text{\# seen with known status})/77774} = \frac{6908}{69083} = 10.0\%.$$

	$\Delta_0 = 0$	$\Delta_0 = 1$	total
$Y_0 = 0$	8491	62175	70666
$Y_0 = 1$	200	6908	7108
total	8691	69083	77774

Table 1: Observed HIV status by CHC/tracking attendance with known status at baseline ($t = 0$)

Numerator - Suppression & HIV-positive: Now our goal is to estimate population-level probability of being suppressed and HIV-positive: $\mathbb{P}(Supp_t^* = 1, Y_t^* = 1)$. This parameter is subject to missing measures on both HIV status and viral load. To simplify notation, let us denote a binary indicator the outcome of interest as $Z_t^* \equiv \mathbb{I}(Supp_t^* = 1, Y_t^* = 1)$. Likewise, let us denote its observed analog as $Z_t \equiv \mathbb{I}(Supp_t = 1, Y_t = 1)$.

Analogous to Section 3.2.4, we formulate the identifiability of the target parameter $\mathbb{P}(Z_t^* = 1)$ as a longitudinal dynamic regime problem. Again, we consider a two component hypothetical intervention on the missingness mechanism:

- $d0$: set $\Delta_t = 1$. In other words, ensure that all individuals attend the CHC/tracking at time t and have known HIV status.
- $d1(Y_t)$: if $(Y_t = 1)$, set $TstVL_t = 1$; else set $TstVL_t = 0$. In other words, if an individual is known to be HIV-positive at time t, ensure that his or her viral load is measured.

Under this hypothetical joint intervention, we would have complete measurement of underlying HIV status and complete measurement of viral load among all HIV-positive individuals.

To identify the probability of suppression under such a hypothetical intervention, suppose we are willing assume

1. HIV status and suppression among those seen at the CHC/tracking and with HIV status measured is representative of HIV status and suppression among those not seen;
2. Among HIV-positive individuals seen at the CHC/tracking, suppression among those with viral load measured is representative of suppression among those missing a viral load measure.

These assumptions are equivalent to the following (sequential) randomization assumptions (Robins, 1986):

$$Z_t^* \perp\!\!\!\perp \Delta_t$$
$$Z_t^* \perp\!\!\!\perp TstVL_t \mid Y_t, \Delta_t = 1$$

Together with the corresponding positivity assumptions, we have

$$\begin{aligned}\mathbb{P}(Z_t^* = 1) &= \sum_{y_t} \mathbb{P}(Z_t = 1 \big| TstVL_t = d1(y_t), Y_t = y_t, \Delta_t = 1)\mathbb{P}(Y_t = y_t | \Delta_t = 1) \\ &= \mathbb{P}(Z_t = 1 | TstVL_t = 1, Y_t = 1, \Delta_t = 1)\mathbb{P}(Y_t = 1 | \Delta_t = 1) \\ &= \mathbb{P}(Supp_t = 1 | TstVL_t = 1)\mathbb{P}(Y_t = 1 | \Delta_t = 1)\end{aligned}$$

In the second equality, we have used that the outcome Z_t is deterministically zero if $Y_t = 0$. In the final equality, we used that the only HIV-positive individuals have their viral loads measured the CHC or tracking. In other words, $TstVL_t = 1$ implies $(Y_t = 1, \Delta_t = 1)$. Therefore, our estimand for the probability of being suppressed and HIV-positive $\mathbb{P}(Supp_t^* = 1, Y_t^* = 1)$ is the proportion of HIV-positive individuals with measured suppression times the prevalence of HIV among those with a known status attending the CHC/tracking.

As shown in Table 2, the first term can be estimated at baseline as

$$\begin{aligned}\hat{\mathbb{P}}(Supp_0 = 1 \mid TstVL_0 = 1) &= \frac{\hat{\mathbb{P}}(Supp_0 = 1, TstVL_0 = 1)}{\hat{\mathbb{P}}(TstVL_0 = 1)} \\ &= \frac{(\#\text{ HIV-positive w/ measured VL} < 500\text{ copies/ml})}{(\#\text{ HIV-positive w/ measured VL})} = \frac{2549}{4983} = 51.2\%.\end{aligned}$$

The second term is our estimand for baseline prevalence. (See above.)

	$TstVL_0 = 0$	$TstVL_0 = 1$	total
$Supp_0 = 0$	2125	2434	4559
$Supp_0 = 1$	0	2549	2549
total	2125	4983	7108

Table 2: Table of viral load testing and observed suppression among known HIV-positive ($Y_0 = 1$) at baseline ($t = 0$).

Proportion - Suppression among HIV-positives: Under these assumptions, we can write the population-level probability of viral suppression among HIV-positive individuals as

$$\begin{aligned}\mathbb{P}(Supp_t^* = 1 \mid Y_t^* = 1) &= \frac{\mathbb{P}(Supp_t^* = 1, Y_t^* = 1)}{\mathbb{P}(Y_t^* = 1)} \\ &= \frac{\mathbb{P}(Supp_t = 1 | TstVL_t = 1)\mathbb{P}(Y_t = 1 | \Delta_t = 1)}{\mathbb{P}(Y_t = 1 | \Delta_t = 1)} \\ &= \mathbb{P}(Supp_t = 1 | TstVL_t = 1)\end{aligned}$$

With this approach, the estimated baseline viral suppression among HIV-positive is 51.2%.

3.6.3 Examining and relaxing assumptions

As discussed above, this approach to estimate HIV prevalence (i.e. the number of HIV-positive individuals in the population) and population-level viral suppression requires on two potentially strong assumptions:

1. HIV prevalence among those attending the CHC/tracking with known status is representative of HIV prevalence among those not attending;
2. For known HIV-positive individuals, suppression among those with a viral load measured at the CHC/tracking is representative of suppression among those without a viral load measured.

However, certain groups may be over-represented and other groups under-represented. If factors affecting CHC/tracking attendance are also associated with underlying HIV status, this will bias our estimates of HIV prevalence. Likewise, if factors affecting viral load measurement are also associated with viral suppression, this will bias our estimates of population-level suppression. In this subsection, we relax these assumptions by assuming they hold within a binary strata. In the following Section 3.6.4, we present our fully adjusted approach, corresponding to the primary analysis.

Denominator - HIV Prevalence: Suppose that we assume within country, HIV prevalence among those testing at the baseline CHC/tracking or attending the baseline CHC/tracking with a documented prior HIV-positive test result ($\Delta_t = 1$) is representative of HIV prevalence among those not ($\Delta_t = 0$). Then we can control for missing HIV tests by estimating HIV prevalence for each country separately and then standardizing to the distribution of these strata in the population:

$$\begin{aligned}\mathbb{P}(Y_t^* = 1) &= \sum_{\text{country}} \mathbb{P}(Y_t = 1 \mid \Delta_t = 1, \text{country})\mathbb{P}(\text{country}) \\ &= \mathbb{P}(Y_t = 1 | \Delta_t = 1, \text{Ugandan})\mathbb{P}(\text{Ugandan}) + \mathbb{P}(Y_t = 1 | \Delta_t = 1, \text{Kenyan})\mathbb{P}(\text{Kenyan}).\end{aligned}$$

With this approach, the estimated baseline prevalence of HIV in SEARCH is

$$\begin{aligned}&\left(\frac{\text{\# HIV-positive Ugandans \& seen with known status}}{\text{\# Ugandans \& seen with known status}}\right)\left(\frac{\text{\# Ugandans}}{\text{\# total pop}}\right) \\ &\quad + \left(\frac{\text{\# HIV-positive Kenyans \& seen with known status}}{\text{\# Kenyans \& seen with known status}}\right)\left(\frac{\text{\# Kenyans}}{\text{\# total pop}}\right) \\ &= \left(\frac{2189}{45033}\right)\left(\frac{50134}{77774}\right) + \left(\frac{4719}{24050}\right)\left(\frac{27640}{77774}\right) = 10.1\%\end{aligned}$$

In this simple scenario, our estimator (G-computation with a saturated parametric regression model) is identical to the inverse probability weighting approach (as used by BCPP in Gaolathe et al. (2016)) and to the targeted maximum likelihood approach (used in this paper).

Numerator - Suppression & HIV-positive: Along with the above assumption on HIV prevalence, suppose that we assume for known HIV-positive individuals and within country, suppression among those with a measured viral load is representative of suppression among those without a measured viral load. Then our estimand is the strata-specific proportion of HIV-positive individuals with measured suppression times the strata-specific prevalence of HIV and then standardized to the distribution of strata in the population:

$$\begin{aligned}
\mathbb{P}(Z_t^* = 1) &= \sum_{\text{country}, y_t} \mathbb{P}(Z_t = 1 \big| TstVL_t = d1(y_t), Y_t = y_t, \Delta_t = 1, \text{country}) \mathbb{P}(Y_t = y_t | \Delta_t = 1, \text{country}) \mathbb{P}(\text{country}) \\
&= \sum_{\text{country}} \mathbb{P}(Z_t = 1 | TstVL_t = 1, Y_t = 1, \Delta_t = 1, \text{country}) \mathbb{P}(Y_t = 1 | \Delta_t = 1, \text{country}) \mathbb{P}(\text{country}) \\
&= \sum_{\text{country}} \mathbb{P}(Supp_t = 1 | TstVL_t = 1, \text{country}) \mathbb{P}(Y_t = 1 | \Delta_t = 1, \text{country}) \mathbb{P}(\text{country})
\end{aligned}$$

Each term can be estimated with the empirical proportion (i.e. contingency tables).

Proportion - Suppression among HIV-positives: To estimate the population-level viral suppression among all HIV-positive adults, we control for missing HIV tests and viral loads by estimating suppression among HIV-positive residents for each country separately and then standardizing to the distribution of residency:

$$\begin{aligned}
\mathbb{P}(Supp_t^* = 1 \mid Y_t^* = 1) &= \sum_{\text{country}} \mathbb{P}(Supp_t^* = 1 \mid Y_t^* = 1, \text{country}) \mathbb{P}(\text{country}) \\
&= \frac{\mathbb{P}(Supp_t = 1 | TstVL_t = 1, \text{Ugandan}) \mathbb{P}(Y_t = 1 | \Delta_t = 1, \text{Ugandan})}{\mathbb{P}(Y_t = 1 | \Delta_t = 1, \text{Ugandan})} \mathbb{P}(\text{Ugandan}) \\
&+ \frac{\mathbb{P}(Supp_t = 1 | TstVL_t = 1, \text{Kenyan}) \mathbb{P}(Y_t = 1 | \Delta_t = 1, \text{Kenyan})}{\mathbb{P}(Y_t = 1 | \Delta_t = 1, \text{Kenyan})} \mathbb{P}(\text{Kenyan}) \\
&= \mathbb{P}(Supp_t = 1 | TstVL_t = 1, \text{Ugandan}) \mathbb{P}(\text{Ugandan}) \\
&+ \mathbb{P}(Supp_t = 1 | TstVL_t = 1, \text{Kenyan}) \mathbb{P}(\text{Kenyan})
\end{aligned}$$

With this approach, the estimated baseline viral suppression among HIV-positive is 48.9%:

$$\begin{aligned}
&\left(\frac{\text{\# HIV-positive Ugandans w/ measured VL} < 500 \text{ copies/ml}}{\text{\# HIV-positive Ugandans w/ measured VL}} \right) \left(\frac{\text{\# Ugandan}}{\text{\# total pop}} \right) \\
&+ \left(\frac{\text{\# HIV-positive Kenyans w/ measured VL} < 500 \text{ copies/ml}}{\text{\# HIV-positive Kenyans w/ measured VL}} \right) \left(\frac{\text{\# Kenyan}}{\text{\# total pop}} \right) \\
&= \left(\frac{643}{1377} \right) \left(\frac{50134}{77774} \right) + \left(\frac{1906}{3606} \right) \left(\frac{27640}{77774} \right) = 48.9\%.
\end{aligned}$$

Again in this simple scenario, our estimator (G-computation with a saturated parametric regression model) is identical to the inverse weighting approach (as used by BCPP in Gaolathe et al. (2016)) and to the targeted maximum likelihood approach (used in this paper).

3.6.4 Primary analysis approach

While stratifying on country weakened our assumptions on missing HIV tests and missing viral loads, they remain quite strong. In practice, there are many potential adjustment variables that could impact the probability of being tested, underlying HIV status, and viral suppression among HIV-positive individuals. Examples include age, sex, occupation, education, socio-economic status, mobility, and community. As the number of potential adjustment variables grow (or we consider continuous variables), the above approach based on contingency tables breaks down due to sparse cells. This problem is further intensified after baseline when the history of testing, care, or suppression could impact current testing, HIV status, and suppression among HIV-positive individuals.

As an alternative, one could use inverse weighting as illustrated in Gaolathe et al. (2016). Targeted maximum likelihood estimation (TMLE), however, is more a efficient and robust approach (van der Laan and Rose, 2011). TMLE combines estimates of expected outcome for potential strata with an estimate of the propensity score (quantification of which types of people more or less represented). As a result, TMLE is a double robust; we will have a consistent estimate if either the strata-specific expected outcome is consistently estimated or the propensity score is consistently estimated. TMLE is also a substitution (plug-in) estimator providing robustness under strong confounding or rare outcomes. Furthermore, rather than rely on parametric regression models (e.g. main terms logistic regression), TMLE further improves robustness by using the machine learning algorithm Super Learner (van der Laan et al., 2007). Super Learner uses cross-validation (i.e. sample splitting) to build the best weighted combination of estimates from a library of candidate algorithms.

Denominator - HIV Prevalence: We refer the reader to Section 3.2.1 for full details on estimating the population-level HIV prevalence in the primary analysis. Our fully adjusted estimate of baseline prevalence is 10.0%.

Numerator - Suppression & HIV-positive: We refer the reader to Section 3.2.4 for full details on estimating the population-level probability of being suppressed and HIV-positive in the primary analysis. Our fully adjusted estimate of the proportion of the population who are HIV-positive and suppressed is 4.5%.

Proportion - Suppression among HIV-positives: In our primary analysis, we estimate the population-level viral suppression among HIV-positive residents by dividing of the estimated probability of being suppressed by the estimated prevalence of HIV (Section 3.2.5). Both the numerator and denominator are fully adjusted for missing tests. With this approach, the estimated baseline viral suppression among HIV-positive is 44.7% - substantially more conservative than the secondary analysis making stronger assumptions about HIV testing and viral load measurement (51.2%).

4 Analysis of the closed cohort of baseline HIV-positive adults

Among enumerated stable residents who are ≥ 15 years of age and known to be HIV-positive at the close of baseline testing ($n = 7108$), we conduct longitudinal analyses evaluating the change in cascade status over time. Specifically, we estimate the probability that such a individual falls into one of six exhaustive and mutually exclusive categories at each time $t = \{0, 1, 2\}$:

1. Dead.
2. Alive and out-migrated.
3. Alive, not out-migrated, and newly diagnosed.
4. Alive, not out-migrated, previously diagnosed, but never on ART.
5. Alive, not out-migrated, with previous ART initiation, but not currently virally suppressed.
6. Alive, not out-migrated, with previous ART initiation, and currently virally suppressed.

Specifically, we estimate the following quantities:

1. **Died**.
$$\mathbb{P}(D_t = 1 \mid Y_0 = 1).$$
By definition of the target population, the probability of being dead at $t = 0$ is 0.

2. **Out-migrated.**
$$\mathbb{P}(M_t = 1, D_t = 0 \mid Y_0 = 1).$$
By definition of the target population, the probability of being out-migrated at $t = 0$ is 0.

3. **New Diagnosis**:

$$\mathbb{P}(pDx_t = 0, M_t = 0, D_t = 0 \mid Y_0 = 1).$$

By definition of the target the population as individuals with known baseline HIV-positive status, the probability of being a new diagnosis for $t > 0$ is 0.

4. **Diagnosed, never on ART**:

$$\mathbb{P}(eART_t = 0, pDx_t = 1, M_t = 0, D_t = 0 \mid Y_0 = 1).$$

5. **Suppression failure:**

$$\mathbb{P}(Supp_t^* = 0, eART_t = 1, pDx_t = 1, M_t = 0, D_t = 0 \mid Y_0 = 1).$$

6. **Suppression success:**

$$\mathbb{P}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, M_t = 0, D_t = 0 \mid Y_0 = 1).$$

The only missingness in these analyses arises from incomplete measurement of viral loads. (Our target population conditions on known baseline HIV-positive status, and death and out-migration are not treated as forms of missingness.) Therefore, the first four quantities can be estimated as empirical proportions. As detailed below, estimators of the probability of suppression failure or success, however, must account for missing viral load measures. As for the open cohort analysis, we use TMLE with Super Learning (with the same library and cross-validation scheme) to adjust for the potentially informative measurement of viral loads. Statistical inference is based on the estimated influence-curve, treating the household as the unit of independence. The corresponding number of HIV-positive individuals in each category is estimated by multiplying the estimated probability times the population size ($n = 7108$).

4.1 Suppression success and failure among baseline known HIV-positive adults

Suppression success: We first focus on identification of the probability of being alive, not out-migrated, previously diagnosed, on ART, and currently virally suppressed. Let

$$Z_t^* = \mathbb{I}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, M_t = 0, D_t = 0).$$

We assume

$$Z_t^* \perp\!\!\!\perp TstVL_t | eART_t, pDx_t, M_t, D_t, \bar{O}_{t-1}, B.$$

This assumption and the resulting estimation scheme can be simplified by noting the following. First, the outcome Z_t^* is deterministically 0 if $D_t = 1$ or $M_t = 1$ or $pDx_t = 0$ or $eART_t = 0$. Furthermore, ever ART use ($eART_t = 1$) implies previous diagnosis ($pDx_t = 1$). Our randomization assumption thus simplifies to

$$Supp_t^* \perp\!\!\!\perp TstVL_t \mid eART_t = 1, M_t = 0, D_t = 0, \bar{O}_{t-1}, B.$$

For individuals who are not dead, not out-migrated and who have previously initiated ART by time t, we assume that within strata defined by baseline demographics, past testing and cascade history (e.g. suppression history), current suppression among individuals who have viral load measured is representative of current suppression among those who are missing viral load measures. We also assume the following positivity assumption:

$$\mathbb{P}(TstVL_t = 1 \big| eART_t = 1, M_t = 0, D_t = 0, \bar{O}_{t-1}, Y_0 = 1, B) > 0.$$

We require individuals who are baseline HIV-positive, not dead, not out-migrated and who have previously initiated ART by time t to have some positive probability of having viral load measured during

CHC/tracking at time t, regardless of covariate values. With these assumptions, our target parameter is identified using the G-computation formula (Robins, 1986):

$$\begin{aligned}
&\mathbb{P}(Z_t^* = 1|Y_0 = 1) = \sum_{l_t} \mathbb{P}(Z_t = 1|TstVL_t = 1, L_t = l_t, Y_0 = 1) \times \mathbb{P}(L_t = l_t|Y_0 = 1) \\
&= \sum_{\bar{o}_{t-1},b} \mathbb{P}(Supp_t = 1|TstVL_t = 1, eART_t = 1, M_t = 0, D_t = 0, \bar{o}_{t-1}, b, Y_0 = 1) \\
&\times \mathbb{P}(eART_t = 1, M_t = 0, D_t = 0, \bar{o}_{t-1}, b \mid Y_0 = 1) \\
&= \sum_{\bar{o}_{t-1},b} \mathbb{P}(Supp_t = 1|TstVL_t = 1, eART_t = 1, M_t = 0, D_t = 0, \bar{o}_{t-1}, b, Y_0 = 1) \\
&\times \mathbb{P}(\bar{o}_{t-1}, b|eART_t = 1, M_t = 0, D_t = 0, Y_0 = 1)\mathbb{P}(eART_t = 1, M_t = 0, D_t = 0|Y_0 = 1).
\end{aligned}$$

where $L_t \equiv (eART_t, pDx_t, M_t, D_t, \bar{O}_{t-1}, B)$. Estimation is based on point treatment TMLE with single intervention node as $TstVL_t$ and adjustment set as $(eART_t, M_t, D_t, \bar{O}_{t-1}, B)$. During estimation, we use knowledge that the joint outcome Z_t is deterministically zero if $eART_t = 0$ OR $D_t = 1$ OR $M_t = 1$.

Suppression failure: Identification and estimation of the probability of suppression failure among individuals known to be HIV-positive at baseline can be implemented analogously. Instead, we estimate this probability by using that these quantities are exhaustive and mutually exclusive:

$$\begin{aligned}
&\mathbb{P}(Supp_t^* = 0, eART_t = 1, pDx_t = 1, M_t = 0, D_t = 0 \mid Y_0 = 1) \\
&= 1 - \mathbb{P}(D_t = 1 \mid Y_0 = 1) - \mathbb{P}(M_t = 1, D_t = 0 \mid Y_0 = 1) - \mathbb{P}(pDx_t = 0, M_t = 0, D_t = 0 \mid Y_0 = 1) \\
&- \mathbb{P}(eART_t = 0, pDx_t = 1, M_t = 0, D_t = 0 \mid Y_0 = 1) \\
&- \mathbb{P}(Supp_t^* = 1, eART_t = 1, pDx_t = 1, M_t = 0, D_t = 0 \mid Y_0 = 1)
\end{aligned}$$

We take the latter approach and use the Delta Method for inference.

4.2 Unadjusted secondary analysis

As a secondary analysis, we control for missing viral load measures by estimating the following 7 probabilities

1. Died: $\mathbb{P}(D_t = 1 \mid Y_0 = 1)$
2. Out-migrated: $\mathbb{P}(M_t = 1, D_t = 0 \mid Y_0 = 1)$
3. New Diagnosis: $\mathbb{P}(pDx_t = 0, M_t = 0, D_t = 0 \mid Y_0 = 1)$
4. Diagnosed, never on ART: $\mathbb{P}(eART_t = 0, pDx_t = 1, M_t = 0, D_t = 0 \mid Y_0 = 1)$
5. On ART, but missed viral load measure:
   $\mathbb{P}(TstVL_t = 0, eART_t = 1, pDx_t = 1, M_t = 0, D_t = 0 \mid Y_0 = 1)$
6. Measured suppression failure:
   $\mathbb{P}(Supp_t = 0, TstVL_t = 1, eART_t = 1, pDx_t = 1, M_t = 0, D_t = 0 \mid Y_0 = 1)$
7. Measured suppression success:
   $\mathbb{P}(Supp_t = 1, TstVL_t = 1, eART_t = 1, pDx_t = 1, M_t = 0, D_t = 0 \mid Y_0 = 1)$

All seven probabilities can be estimated with empirical proportions. For example, the probability of dying is estimated as the number of baseline HIV-positive subjects who died by time t divided by the population size ($n = 7108$).

4.3 Stratified analysis

We implement the same estimators, but now within the following strata defined by baseline variables included in B.

1. Sex (Male vs. Female)
2. Age (Younger: 15-24 years vs. Older: > 24 years) at baseline
3. Country of residence (Uganda vs. Kenya)

5 Closed cohort analyses treating death and outmigration as right-censoring events

In the previous section, death and out-migration are treated as outcomes. As an alternative, we conduct complimentary longitudinal analyses with right-censoring at death or out-migration. These analyses adjust for the potentially informative nature of this censoring in addition to informative missingness of viral load measures where applicable. First, in a closed cohort of baseline enumerated, stable adult residents without an HIV diagnosis prior to the baseline testing round, we estimate the proportion who are tested at least once by the close of two and three rounds of testing ($t = \{1, 2\}$). Second, among adults diagnosed as HIV-positive by the close of baseline testing, we estimate the proportion who have initiated ART by $t = \{1, 2\}$. Third, among the cohort of adults diagnosed as HIV-positive by the close of baseline testing, we estimate the proportion virally suppressed at $t = \{1, 2\}$, while accounting for missing viral load measures. We further evaluate evolution in post-baseline viral suppression within subgroups defined by baseline cascade status (prior diagnosis, ART use, and viral suppression). Finally, we consider univariate and multivariate predictors of these outcomes. Adjustment employs TMLE with Super Learning (as described previously).

5.1 Target population and outcomes of interest

We consider the following outcomes.

1. $ever.test_t$: indicator that an individual has ever had an HIV test (i.e. any $TstHIV_j = 1$, $j \leq t$). This is a counting process that jumps at most once (i.e. if $ever.test_j = 1$, then $ever.test_t = 1$ deterministically for $t > j$). If an individual has no record of having tested, he or she is assumed not to have tested.

2. $eART_t$: indicator that an individual has ever been on ART. This is a counting process that jumps at most once (i.e. if $eART_j = 1$, then $eART_t = 1$ deterministically for $t > j$). If an individual has no record of having initiated ART, he or she is assumed not to have initiated ART.

3. $Supp_t^*$: indicator that an HIV-positive individual has achieved viral suppression at year t CHC/tracking (i.e. if $VLV_t^* < 500$ copies/ml then $Supp_t^* = 1$). This variable is not a counting process and is measured only if $TstVL_t = 1$ (i.e. $Supp_t = Supp_t^* \times TstVL_t$).

We define the following outcome-specific target populations:

- We restrict the target population for all outcomes to baseline enumerated stable, adult (≥ 15 years of age) residents.
- For outcome $ever.test_t$, we further restrict the target population to exclude those with a prior diagnosis at baseline ($pDx_0 = 1$).
- For outcomes $(eART_t, Supp_t^*)$, we restrict the target population to known baseline HIV-positive individuals ($Y_0 = 1$).

5.2 Target parameters

5.2.1 Cascade probabilities over time

We consider the following hypothetical interventions of interest. For all outcomes, we aim to evaluate the counterfactual proportion that would have been observed in the absence of death/out-migration. In other words, we intervene to set $D_t = 0$ and $M_t = 0$. For suppression outcome $Supp_t^*$, we further intervene to set $TstVL_t = 1$ to ensure that viral load is measured at the CHC/tracking at time t. We generically denote these interventions of interest as d.

For each outcome, we aim to estimate its time point-specific mean under an intervention to prevent censoring and for suppression, to ensure viral load measurement. Let Z_t refer generically to each outcome above, and let $Z_t(d)$ refer to the counterfactual value of this outcome under regime d. We consider the following target parameters:

- For all outcomes, we estimate the intervention-specific mean outcome $\mathbb{E}[Z_t(d)]$ for $t = \{1, 2\}$.
- Additionally for suppression outcome $Supp_t^*$, we estimate the post-baseline ($t = \{1, 2\}$) probability of suppression within the following subgroups defined by baseline cascade status.
  - **Baseline new diagnoses:** Among baseline new diagnoses (i.e. those with no record of prior care at baseline), we estimate the overall post-baseline suppression probability

    $$\mathbb{E}[Supp_t^*(d)|pDx_0 = 0].$$

  - **Baseline prior diagnosis without ART:** Among individuals with a record of prior diagnosis but no record of prior ART initiation at baseline, we estimate the overall post-baseline suppression probability

    $$\mathbb{E}[Supp_t^*(d)|pDx_0 = 1, eART_0 = 0].$$

  - **Prior ART initiation:** Among individuals with a record of ART initiation prior to baseline, we estimate the overall post-baseline suppression probability

    $$\mathbb{E}[Supp_t^*(d)|eART_0 = 1].$$

    We further estimate post-baseline suppression probability within subgroups defined by viral suppression at baseline (measured suppression success, measured suppression failure, or missing baseline viral load).

5.2.2 Predictors of cascade failures

For all outcomes at $t = 2$, we estimate variable importance measures on an absolute scale (statistical analog of the causal risk difference), treating each variable in B in turn as the intervention variable, and the remainder as the adjustment set. We consider hypothetical interventions as above to prevent censoring for all outcomes; for $Supp_t^*$ we consider the additional hypothetical intervention to ensure that viral load is measured ($TstVL_t = 1$).

5.3 Identification

In the primary analyses, we assume that within strata defined by past testing, cascade history, and baseline covariates, those who remain alive and resident in the community are representative of those who are censored with respect to each of the outcomes considered.

$$Z_t(d) \perp\!\!\!\perp C_j \mid B, \bar{O}_{j-1}, C_{j-1} = 0, \quad \text{for } j \leq t \text{ and } t = \{1, 2\}.$$

We further assume positivity; there are no strata (defined by testing history, cascade history and baseline demographics) in which all individuals are censored:

$$\mathbb{P}(C_j = 0|B, \bar{O}_{j-1}, C_{j-1} = 0) > 0, \quad \text{for } j \leq t.$$

For the suppression outcome $Supp_t^*$, we also assume that within strata defined by past testing, cascade history (including current ART use), and baseline demographics, suppression among individuals with a viral load measured at CHC/tracking is representative of suppression among those missing a viral load measurement:

$$Z_t(d) \perp\!\!\!\perp TstVL_t \mid B, \bar{O}_{j-1}, eART_t, C_t = 0, \quad \text{for } t = \{1, 2\}.$$

Finally, we assume that among individuals who remain alive and resident in the community, there are no substrata (defined by baseline covariates, past testing and cascade history) in which all individuals are missing a viral load test:

$$\mathbb{P}(TstVL_t = 1 \mid B, \bar{O}_{t-1}, eART_t, C_t = 0) > 0.$$

5.4 Estimators

- For target parameters corresponding to the intervention-specific mean (overall and within the pre-specified subgroups), we implement the following estimators.
  - Adjusted estimators, which control for informative censoring (and informative missing viral load for the suppression outcome) using longitudinal TMLE. Adjustment sets are implied by independence assumptions above. Nuisance parameters are estimated using Super Learner, with 5-fold cross-validation and library as specified in the open cohort analysis.
  - Unadjusted estimators correspond to simple empirical proportions among uncensored individuals (i.e. conditioning on $C_t = 0$) as well as measured viral loads ($TstVL_t = 1$) for the suppression outcome.
- We estimate univariate and multivariate predictors of each outcome an absolute risk scale (corresponding to unadjusted and adjusted risk differences, respectively).
  - We consider each of the following variables included in the baseline characteristics B in turn as the predictor of interest:
    - Age: 4-level categorical variable
    - Sex
    - Mobility: > 1 mo/past year away from community or not
    - Marital Status: ever vs. never married
    - Education: less than primary, primary, secondary or higher
    - Occupation: formal sector, high risk informal sector, low risk informal sector, other, no job/disabled
    - Household wealth quintile based on a principal components analysis of a baseline household consumption survey (Chamie et al., 2016).
  - Using longitudinal TMLE, multivariate predictors adjust for the remaining variables in B, informative censoring, and informative missing viral load measures for the suppression outcome. Again, adjustment sets for censoring and viral load testing are implied by independence assumptions above. Nuisance parameters are estimated using Super Learner, with 5 fold cross-validation and library as specified in the open cohort analysis.
  - Univariate predictors are evaluated conditional on being uncensored ($C_t = 0$), as well as conditional on measured viral load ($TstVL_t = 1$) for suppression outcome, and without adjusting for other variables in B.

  Statistical inference is based on the estimated influence-curve, treating the household as the unit of independence.

6 Acknowledgements:

Research reported in this presentation was supported by Division of AIDS, NIAID of the National Institutes of Health under award numbers (U01AI099959, R37AI051164, R01-AI074345) and in part by the President's Emergency Plan for AIDS Relief, Bill and Melinda Gates Foundation, and Gilead Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH, PEPFAR, Bill and Melinda Gates Foundation, or Gilead. The SEARCH project gratefully acknowledges the Ministries of Health of Uganda and Kenya, our research team, collaborators and advisory boards, and especially all communities and participants involved.

References

G. Chamie, T.D. Clark, J. Kabami, K. Kadede, E. Ssemmondo, R. Steinfeld, G. Lavoy, D. Kwarisiima, N. Sang, V. Jain, H. Thirumurthy, T. Liegler, L. Balzer, et al. A hybrid mobile HIV testing approach for population-wide HIV testing in rural East Africa. *Lancet HIV*, January, 2016.

T. Gaolathe, K.E. Wirth, M.P. Holme, J. Makhema, S. Moyo, et al. Botswana's progress toward achieving the 2020 UNAIDS 90-90-90 antiretroviral therapy and virological suppression goals: a population-based survey. *Lancet HIV*, Online, 2016. doi: 10.1016/S2352-3018(16)00037-0.

M.A. Hernán, E. Lanoy, D. Costagliola, and J.M. Robins. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & Clinical Pharmacology & Toxicology*, 98(3):237–242, 2006.

E. Polley and M. van der Laan. *SuperLearner: Super Learner Prediction*, 2014. URL `{http://CRAN.R-project.org/package=SuperLearner}`. R package version 2.0-15.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL `http://www.R-project.org`.

J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods–application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986. doi: 10.1016/0270-0255(86)90088-6.

J.M. Robins, L. Orellana, and A. Rotnitzky. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27(23):4678–4721, 2008.

Joshua Schwab, Samuel Lendle, Maya Petersen, and Mark van der Laan. *ltmle: Longitudinal Targeted Maximum Likelihood Estimation*, 2016. URL `http://CRAN.R-project.org/package=ltmle`. R package version 0.9-9.

UNAIDS. 90-90-90 an ambitious treatment target to help end the AIDS epidemic, 2014. URL `http://www.unaids.org/en/resources/documents/2014/90-90-90`.

M. van der Laan and S. Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York Dordrecht Heidelberg London, 2011.

M.J. van der Laan and M.L. Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1):Article 3, 2007.

M.J. van der Laan and D.B. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):Article 11, 2006. doi: 10.2202/1557-4679.1043.

M.J. van der Laan, E.C. Polley, and A.E. Hubbard. Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1):25, 2007. doi: 10.2202/1544-6115.1309.