Supplementary Materials: Far from MCAR: obtaining population-level estimates of HIV viral suppression

February 25, 2020

1 eAppendix1: Further details on the identifiability results

Recall the following variable definitions:

- t: time of annual community-wide testing
- HIV_t^*: possibly unmeasured indicator of HIV-positive serostatus at time t
- Δ_t^{HIV}: indicator that an individual was seen at community-wide testing and had "known" HIV status at time t - due to a negative test result at time t, or a positive result at or before time t
- HIV_t: observed HIV status, defined as $HIV_t = \Delta_t^{HIV} \times HIV_t^*$
- $Supp_t^*$: possibly unmeasured indicator of HIV viral suppression (<500cps/mL) at time t
- Δ_t^{Supp}: an indicator of viral load measurement at time t
- $Supp_t$: observed viral suppression status, defined as $Supp_t = \Delta_t^{Supp} \times Supp_t^*$

Our goal is to estimate the population-level HIV viral suppression at time t:

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

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. For ease of notation, let $Z_t^* \equiv \mathbb{I}(Supp_t^* = 1, HIV_t^* = 1)$; then we can define our target of interest as

$$\frac{\mathbb{P}(Z_t^* = 1)}{\mathbb{P}(HIV_t^* = 1)}.$$

We first focus on identifying the denominator and then the numerator. By taking the ratio of the numerator to the denominator, we can combine these estimates to obtain an estimate of population-level viral suppression among HIV-positive adults. Throughout, we use that the expectation of a binary variable is equivalent to the probability of event. Suppose, for example, $X = \{0, 1\}$; then $\mathbb{E}(X) \;=\; \mathbb{P}(X = 1)$.

1.1 Denominator - HIV Prevalence $\mathbb{P}(HIV_t^* = 1)$

1.1.1 Unadjusted

Assuming $HIV_t^* \perp\!\!\!\perp \Delta_t^{HIV}$, we have the following identifiability result:

$$\begin{aligned}\mathbb{P}(HIV_t^* = 1) &= \mathbb{P}(HIV_t^* = 1 \mid \Delta_t^{HIV} = 1) \\ &= \mathbb{P}(HIV_t = 1 \mid \Delta_t^{HIV} = 1)\end{aligned}$$

The first equality holds under our identifiability assumption: $HIV_t^* \perp\!\!\!\perp \Delta_t^{HIV}$. The second equality holds by definition of the outcome: $HIV_t = \Delta_t^{HIV} \times HIV_t^*$. The identifiability result can equivalently be expressed $\mathbb{E}(HIV_t \mid \Delta_t^{HIV} = 1)$.

1.1.2 Baseline adjustment

We can relax the above assumption by stratifying on baseline covariates B. Along with the corresponding positivity assumption, suppose we are willing assume: $HIV_t^* \perp\!\!\!\perp \Delta_t^{HIV} \mid B$. Then we obtain the following identifiability result:

$$\begin{aligned}\mathbb{P}(HIV_t^* = 1) &= \sum_b \Big[\mathbb{P}(HIV_t^* = 1 \mid B = b) \times \mathbb{P}(B = b)\Big] \\ &= \sum_b \Big[\mathbb{P}(HIV_t^* = 1 \mid \Delta_t^{HIV} = 1, B = b) \times \mathbb{P}(B = b)\Big] \\ &= \sum_b \Big[\mathbb{P}(HIV = 1 \mid \Delta_t^{HIV} = 1, B = b) \times \mathbb{P}(B = b)\Big]\end{aligned}$$

The identifiability result can equivalently be expressed $\mathbb{E}\big[\mathbb{E}(HIV_t \mid \Delta_t^{HIV} = 1, B)\big]$.

1.2 Time-varying adjustment

Let L_t denote the full set of baseline demographics and prior HIV testing behavior (e.g. number and location). Assuming positivity and $HIV_t^* \perp\!\!\!\perp \Delta_t^{HIV} \mid L_t, HIV_{t-1} = 0$, we have the following identifiability result:

$$\begin{aligned}&\mathbb{P}(HIV_t^* = 1) \\ &\quad= \mathbb{P}(HIV_t^* = 1 \mid HIV_{t-1} = 1)\mathbb{P}(HIV_{t-1} = 1) + \mathbb{P}(HIV_t^* = 1 \mid HIV_{t-1} = 0)\mathbb{P}(HIV_{t-1} = 0) \\ &\quad= \mathbb{P}(HIV_{t-1} = 1) + \mathbb{P}(HIV_t^* = 1 \mid HIV_{t-1} = 0)\mathbb{P}(HIV_{t-1} = 0) \\ &\quad= \mathbb{P}(HIV_{t-1} = 1) + \mathbb{P}(HIV_{t-1} = 0) \sum_{l_t} \Big[\mathbb{P}(HIV_t^* = 1 \mid L_t = l_t, HIV_{t-1} = 0)\mathbb{P}(L_t = l_t \mid HIV_{t-1} = 0)\Big] \\ &\quad= \mathbb{P}(HIV_{t-1} = 1) + \mathbb{P}(HIV_{t-1} = 0) \sum_{l_t} \Big[\mathbb{P}(HIV_t = 1 \mid \Delta_t^{HIV} = 1, L_t = l_t, HIV_{t-1} = 0)\mathbb{P}(L_t = l_t \mid HIV_{t-1} = 0)\Big].\end{aligned}$$

The first equality is by the Law of Total Probability and the definition of conditional probabilities. In the second equality, we use that individuals known to be HIV-positive at the previous time-point (i.e. with $HIV_{t-1} = 1$) are HIV-positive at the next time-point ($HIV_t^* = 1$).

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 summation over adjustment variables l_t generalizes to an integral for continuous-valued variables:

$$\mathbb{P}(HIV_{t-1} = 1) + \mathbb{P}(HIV_{t-1} = 0) \times \mathbb{E}\Big[\mathbb{E}(HIV_t \mid \Delta_t^{HIV} = 1, L_t, HIV_{t-1} = 0)\Big| HIV_{t-1} = 0\Big]$$

1.3 Numerator - Probability of Suppression & HIV-positivity: $\mathbb{P}(Z_t^* = 1)$

This parameter is subject to missing measures on both HIV status and viral loads.

1.3.1 Unadjusted

We formulate the identifiability of the target parameter $\mathbb{P}(Z_t^* = 1)$ as a longitudinal dynamic regime problem, corresponding to the following hypothetical interventions (e.g. Hernán et al. (2006); van der Laan and Petersen (2007); Robins et al. (2008)):

- $d0$: set $\Delta_t^{HIV} = 1$.
- $d1(HIV_t)$: if $(HIV_t = 1)$, set $\Delta_t^{Supp} = 1$; else, set $\Delta_t^{Supp} = 0$.

We note that other approaches are also possible, and several alternative formulations result in the same statistical estimand.

Along with the corresponding positivity assumptions, suppose we are willing assume (Robins, 1986):

$$Z_t^* \perp\!\!\!\perp \Delta_t^{HIV}$$
$$Z_t^* \perp\!\!\!\perp \Delta_t^{Supp} \mid HIV_t, \Delta_t^{HIV} = 1$$

Then, we have

$$\begin{aligned}\mathbb{P}(Z_t^* = 1) &= \sum_{hiv_t} \left[\mathbb{P}(Z_t^* = 1 \mid \Delta_t^{Supp} = d1(hiv_t), hiv_t, \Delta_t^{HIV} = 1) \times \mathbb{P}(hiv_t = 1 \mid \Delta_t^{HIV} = 1) \right] \\ &= \mathbb{P}(Z_t^* = 1 \mid \Delta_t^{Supp} = 1, HIV_t = 1, \Delta_t^{HIV} = 1) \times \mathbb{P}(HIV_t = 1 \mid \Delta_t^{HIV} = 1) \\ &= \mathbb{P}(Supp_t = 1 \mid \Delta_t^{Supp} = 1, HIV_t = 1, \Delta_t^{HIV} = 1) \times \mathbb{P}(HIV_t = 1 \mid \Delta_t^{HIV} = 1)\end{aligned}$$

In the second equality, we use that Z_t^* is deterministically zero for HIV-negative persons (i.e., those with $HIV_t = 0$ and $\Delta_t^{HIV} = 1$). Also in the second equality, we use our definition of dynamic treatment rule $d1(HIV_t)$. In the third equality, we use that $Z_t^* = Supp_t$ for persons who are known to be HIV-positive and have their viral load measured. Since $HIV_t = 1$ implies $\Delta_t^{HIV} = 1$, we can simplify our estimand to

$$\mathbb{P}(Supp_t = 1 \mid \Delta_t^{Supp} = 1, HIV_t = 1) \times \mathbb{P}(HIV_t = 1 \mid \Delta_t^{HIV} = 1)$$

1.3.2 Baseline adjustment

Following the same approach, we can relax the above assumptions by stratifying on baseline covariates B. Along with the corresponding positivity assumptions, suppose we are willing assume (Robins, 1986):

$$Z_t^* \perp\!\!\!\perp \Delta_t^{HIV} \mid B$$
$$Z_t^* \perp\!\!\!\perp \Delta_t^{Supp} \mid HIV_t, \Delta_t^{HIV} = 1, B$$

Then we would obtain the following identifiability result; the steps are analogous to the above and thus omitted:

$$\mathbb{P}(Z_t^* = 1) = \sum_b \left[\mathbb{P}(Supp_t = 1 \mid \Delta_t^{Supp} = 1, HIV_t = 1, B = b) \times \mathbb{P}(HIV_t =\mid \Delta_t^{HIV} = 1, B = b) \times \mathbb{P}(B = b) \right]$$

1.4 Time-varying adjustment

Let ART_t be an indicator of ART initiation by time t, and let X_t denote the full set of baseline and time-updated covariates. We could follow an analogous approach to further relax the identifiability assumptions

by adjusting for ART_t and X_t. Instead, we make use of the fact that UNAIDS target is focused on ART-induced suppression and define $Supp_t^* = 0$ if $ART_t = 0$ (Joint United Nations Programme on HIV/AIDS (UNAIDS), 2014). Along with the corresponding positivity assumption, we assume (Robins, 1986):

$$Z_t^* \perp\!\!\!\perp \Delta_t^{Supp} \mid ART_t = 1, X_t$$

Then we have the following identifiability result:

$$\begin{aligned}\mathbb{P}(Z_t^* = 1) &= \sum_{x_t} \left[\mathbb{P}(Supp_t = 1 \mid \Delta_t^{Supp} = 1, ART_t = 1, X_t = x_t) \times \mathbb{P}(ART_t = 1, X_t = x_t) \right] \\ &= \mathbb{P}(ART_t = 1) \times \sum_{x_t} \left[\mathbb{P}(Supp_t = 1 \mid \Delta_t^{Supp} = 1, ART = 1, X_t = x_t) \times \mathbb{P}(X_t = x_t \mid ART_t = 1) \right]\end{aligned}$$

In the second equality, we used the definition of conditional probability to rewrite the joint probability $\mathbb{P}(ART_t = 1, X_t = x_t)$ as $\mathbb{P}(X_t = x_t \mid ART_t = 1) \times \mathbb{P}(ART_t = 1)$. 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. As before, we can rewrite the estimand in terms of expectations, which is more appropriate for continuous-valued adjustment variables:

$$\mathbb{P}(ART_t = 1) \times \mathbb{E}\left[\mathbb{E}(Supp_t \mid \Delta_t^{Supp} = 1, ART = 1, X_t) \middle| ART_t = 1 \right]$$

2 eAppendix2: Step-by-step implementation to adjust for missing outcomes

Implementation for estimating population-level viral suppression is given in the next section. This section, instead, considers various approaches to estimation in a generic setting where outcome data are missing.

If we were willing to assume outcomes were missing-completely-at-random (MCAR), then we could take an unadjusted approach: simply count of the number with the outcome and divide by the number with measured. This is equivalent to taking the empirical proportion among those with measured outcomes.

If we were willing to assume discrete-valued baseline covariates were the only common causes of measurement and health outcomes, then we could implement a **non-parametric, stratification-based approach**. Within each level of adjustment strata, we would take the number with the outcome and divide by the number measured, Finally, we would average (i.e., marginalize) the strata-specific estimates. This approach, however, quickly breaks down when there is a moderately sized adjustment set or a single continuous variable. Suppose, for example, we wanted to adjust for age deciles; then we would have 1024 strata in which we needed to count the number with the outcome and divide by the number measured.

Alternative approaches allow us to smooth over strata with weak support. To adjust for missing data on the outcome, we could implement **parametric G-computation** as follows.

1. Among those measured, regress the outcome on the adjustment covariates. For example, we may be willing to assume the probability of suppression (among those measured) is accurately described by a logistic regression model with main terms for the adjustment factors.
2. Use the resulting coefficients from #1 to predict the outcome for **all** observations.
3. Average the predictions.

To adjust for missing data on the outcome, we could also take an **inverse-weighting** approach:

1. Regress the measurement indicator on the adjustment covariates. For example, we may be willing to assume the probability of viral load testing is accurately described by a logistic regression model with main terms for the adjustment factors.

2. Use the resulting coefficients from #1 to predict the probability of measurement for all observations (a.k.a. "the propensity score").
3. Calculate the weights: an indicator of measurement, divided by predicted probability of being measured.
4. Average the weighted outcomes.

It is important to note that if saturated regressions are used to adjust for the same variables, then parametric G-computation and inverse-weighting are equivalent to each other and also to the non-parametric, stratification based approach.

To flexibly adjust for a larger set of baseline and time-varying covariates, we could implement **targeted maximum likelihood estimation (TMLE) with Super Learner**. Briefly, we would

1. Among those measured, use Super Learner to flexibly model the relationship between the outcome and adjustment factors.
2. Use the output from #1 to predict the outcome for all observations.
3. Target these machine learning-based predictions with information in the propensity score (i.e. probability of measurement, given the adjustment set), which was also fit with Super Learner.
4. Average the targeted predictions

3 eAppendix3: Further details on estimator implementation

In the following, we provide additional details on the estimation algorithms applied to the real data. Complete `R` code, sufficient to replicate these analyses, is available at
`https://github.com/LauraBalzer/Far-From-MCAR`.

3.1 Unadjusted

The unadjusted estimator is simply the number with suppressed viral replication divided by the number with viral load measurement. This point estimate with statistical inference can be obtained via the `ltmle` package (Schwab et al., 2017) using the `id` option to specify the community as the unit of independence.

3.2 Baseline adjustment

To estimate both the numerator and denominator of the baseline adjusted parameter, we implemented a non-parametric stratification approach to control for mutually exclusive-and-exhaustive strata, defined by age group (15-19yrs, 20-29yrs, 30-39yrs, 40-49yrs, 50-59yrs, 60+yrs), sex, and community. To do so, we implemented inverse-weighting with the propensity score (i.e. the probability of measurement) estimated using a saturated logistic regressions. We again used the `ltmle` package with the `id` option to specify the community as the unit of independence. Since we used a saturated regression, identical results would have been obtained with parametric G-computation (using saturated regressions). We obtained inference for population-level viral suppression (i.e. the ratio of the numerator to the denominator) with the Delta Method, again treating the community as the unit of independence.

3.3 Time-varying adjustment

To estimate both the numerator and denominator of the fully adjusted parameter, we implemented TMLE with Super Learner within each community separately. The 32 baseline demographic variables were selected based on known HIV epidemiology and included age, sex, marital status, education, occupation, alcohol use, wealth index, and measures of mobility. When estimating HIV prevalence, we also adjusted for baseline HIV testing behavior and incorporated deterministic knowledge on HIV status (i.e. once a person is HIV positive, they remain HIV positive). When estimating the probability of being HIV-infected and

suppressed, we additionally adjusted for baseline HIV testing behavior and prior viral suppression as well as incorporated deterministic knowledge on ART use (i.e. persons not on ART are not suppressed). We used the same Super Learner library when estimating the outcome regressions and the propensity scores:

- Logistic regression after screening based on 10 highest correlations
- Logistic regression after screening based on significant correlations ($p < 0.1$)
- Generalized additive model after screening based on 10 highest correlations
- Generalized additive model after screening based on significant correlations ($p < 0.1$)
- The mean (i.e. unadjusted)

We selected this library for ease of interpretation, reduced computational burden, and to avoid over-fitting. In sensitivity analysis, using a larger library yielded nearly identical results. As before, we implemented with the `ltmle` package, which, by default, truncates the estimated propensity scores at (0.01, 1). We then combined the community-specific estimates together and again used the Delta Method to obtain inference for population-level viral suppression (i.e. the ratio of the numerator to the denominator).

4 eTable1

eTable1: Baseline characteristics of the adult resident population – overall and by measurement and outcome status. Each column is a subset of the former. Metrics are in N (%).

	Enumerated adult population (N=79,818)	HIV serostatus known (N=71,402)	HIV-positive serostatus (N=7009)	Viral load measured (N=5332)
Region				
Western Uganda	25305 (32)	23383 (33)	1469 (21)	1076 (20)
Eastern Uganda	26162 (33)	23500 (33)	762 (11)	586 (11)
Kenya	28351 (36)	24519 (34)	4778 (68)	3670 (69)
Sex				
Female	43770 (55)	40091 (56)	4659 (66)	3599 (67)
Male	36048 (45)	31311 (44)	2350 (34)	1733 (33)
Age category				
15-19 yrs	16991 (21)	15080 (21)	229 (3)	167 (3)
20-29 yrs	22220 (28)	19649 (28)	1865 (27)	1378 (26)
30-39 yrs	14884 (19)	13357 (19)	2261 (32)	1721 (32)
40-49 yrs	10005 (13)	9079 (13)	1506 (21)	1160 (22)
50-59 yrs	7011 (9)	6333 (9)	764 (11)	616 (12)
$\geq$ 60 yrs	8707 (11)	7904 (11)	384 (5)	290 (5)
Marital status[a]				
Single	23692 (30)	20314 (28)	501 (7)	352 (7)
Married	46684 (58)	42417 (59)	4813 (69)	3647 (68)
Widowed/divorced/separated	9228 (12)	8465 (12)	1688 (24)	1327 (25)
Education level[b]				
Below primary	50912 (64)	46413 (65)	5258 (75)	4067 (76)
Completed primary	11478 (14)	10095 (14)	814 (12)	602 (11)
Any secondary or higher	17266 (22)	14759 (21)	921 (13)	650 (12)
Occupation[c]				
Formal	19753 (25)	16884 (24)	362 (5)	257 (5)
High-risk informal	3235 (4)	2762 (4)	678 (10)	383 (7)
Low-risk informal	48753 (61)	44738 (63)	5193 (74)	4133 (78)
Other	3864 (5)	3308 (5)	387 (6)	288 (5)
No job	3992 (5)	3500 (5)	384 (5)	266 (5)
Household wealth quintile[d]				
First, least wealth	12078 (15)	10952 (15)	1109 (16)	827 (16)
Second	13474 (17)	12188 (17)	1147 (16)	914 (17)
Third	15448 (19)	13932 (20)	1260 (18)	979 (18)
Fourth	17384 (22)	15507 (22)	1535 (22)	1184 (22)
Fifth, most wealth	21235 (27)	18636 (26)	1854 (26)	1392 (26)

[a]Marital status missing for N=214 (0.3%); [b]Education level missing for n=162 (0.2%); [c]Formal sector was a teacher, student, government worker, military worker, health worker, or factory worker. High-risk informal was a fishmonger, fisher, bar owner, bar worker, transportation worker, or tourism worker. Low-risk informal was a farmer, shopkeeper, market vendor, hotel worker, homemaker, household worker, construction worker, or miner. Occupation missing for N=221 (0.3%). [d]Household wealth quintile determined through principle component analysis; missing for N=199 (0.2%).

References

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.

Joint United Nations Programme on HIV/AIDS (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`.

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.

J. Schwab, S. Lendle, M. Petersen, and M van der Laan. *ltmle: Longitudinal Targeted Maximum Likelihood Estimation*, 2017. URL `http://CRAN.R-project.org/package=ltmle`.

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.