P3 Statistical Analysis Plan

Study Title	P3 (Prepared, Protected, emPowered)
Principal Investigator	Lisa Hightow, MD, MPH Professor, Division of Infectious Diseases Behavior and Technology Lab at UNC iTech UNC School of Medicine Chapel Hill, NC
ClinicalTrials.gov Identifier	NCT03320512
Authors	Lindsey Schader, MS PhD Candidate Biostatistics Department Laney Graduate School Atlanta, GA David Benkeser, PhD Assistant Professor Biostatistics Department Rollins School of Public Health Atlanta, GA
Date	July 21, 2022
Version	2

1 Revision History

Version	Date	Summary of Changes
Final	05-JAN-2022	
Amendment 1	07-JULY-2022	**Minor Changes Throughout**: Updates to notation to maintain consistency. Also added/removed and changed sentences as necessary to improve the organization of the protocol. **Section 3.1 (see also Table 4, 8, and 10)**
	04-MARCH-2022	Changed the primary study objective to consider FTC-tp and TFV-dp levels in the blood separately. **Section 3.2 (see also Tables 5, 6, 8, and 12)**
	28-JAN-2022	Removed secondary objective II, where FTC-tp and TFV-dp were analyzed as continuous outcomes.
	13-APR-2022	Added self-reported weekly and monthly PrEP use as outcomes of interest for the secondary study objective.
	10-JAN-2022	Added exploratory outcomes as determined by the study team. **Section 3.3 (see Table 6)**
	10-JULY-2022	Updated the cost effectiveness objective to reflect that it is a comparison between P3 and P3+ as originally defined in the protocol. **Section 4.2 (see Table 7)**
	09-MARCH-2022	Updated super learner libraries. We added a simple model and regularization models (glment and stepwise selection) out of concern regarding the large number of confounders in our models.

	26-JUNE-2022	Removed statement regarding double robustness of TMLE.
	13-APR-2022	Added race/ethnicity to the variable list.
		Section 4.3 (see also Table 8)
	13-APRIL-2022	Hypothesis testing for 3 and 6 months was separated into two different hypothesis tests as opposed to considering a joint test of no treatment effect at both 3 and 6 months. This effects the primary and secondary/exploratory objectives.
		Section 4.4
	13-JULY-2022	Added this section to explain cost analysis.
		Section 6.1
	25-JUNE-2022	Timeline updated to reflect realized dates
		Section 6.2
	16-MARCH-2022	Updated the imputation steps for missing survey data.
		Section 6.3
	13-APR-2022	Race/ethnicity was added as a confounder to control for in all analyses (see also Table 7).
	21-JUNE-2022	Removed longitudinal TMLE as a sensitivity analysis for unmeasured, time-varying confounding.
		Section 6.4
	09-MARCH-2022	Added a sensitivity analysis for the primary objective.
		Supplement
	21-JUNE-2022	Rewrote the supplement to provide adequate detail regarding the TMLE procedure for incorporating survey data into the analysis of the primary objective.

Contents

Table 4: Primary outcome measures of interest.

Outcome	Definition(s)	Collection Method
Protective level of TFV-dp in the blood	TFV-dp levels in the blood consistent with PrEP use >4 days per week	DBS
Protective level of FTC-tp in the blood	FTC-tp levels in the blood consistent with PrEP use >4 days per week	DBS

2 Abbreviations and Definitions

CASI - Computer-assisted self-interviewing
DBS - Dry blood spot
SOC - Standard of care
STI - Sexually transmitted infection
TMLE - Doubly robust targeted minimum-loss based estimation

3 Study Objectives

3.1 Primary Study Objectives

The primary objective of this analysis is to assess the efficacy of P3/P3+ at improving PrEP adherence, as compared to Standard of Care (SOC), at both 3 months and 6 months of follow-up. The primary outcomes of interest will be a protective level of TFV-dp and FTC-tp in dry blood spot (DBS) plasma. These two measures will be analyzed separately for the primary objective.

3.2 Secondary Study Objectives

As detailed in the study protocol, the secondary objectives of this analysis are as follows:

1. Replicate the primary objectives with P3 and P3+ considered separately against the SOC group (i.e. P3 vs. SOC and P3+ vs SOC).

2. Analyze the effect of the intervention on self-reported weekly PrEP use and self-reported monthly PrEP use.

3. Analyze the effect of intervention on a panel of self-reported outcomes which include, PrEP retention and persistence, PrEP clinical care, sexual risk, and incident sexually transmitted infection (STI).

The outcomes for secondary objective one are the same as the primary outcome and are listed in table 4. Table 5 lists outcomes for secondary objectives two and three. These outcomes will be assessed at both the 3 month time point and the 6 month time point.

3.3 Cost Effectiveness Objectives

The last objective of the study is to assess the cost effectiveness of P3+ compared to P3. We will compare cost effectiveness of interventions by calculating $\Delta C/\Delta E$, where ΔC is the difference in mean costs of the intervention and ΔE is the mean difference in the effectiveness.

Table 5: Secondary Objective Outcomes

Outcome	Definition	Collection Method
TFV-dp	Same as primary objective	DBS
FTC-tp	Same as primary objective	DBS
Weekly PrEP Use	Self-reported number of days in the past week that the patient took their PrEP medication	CASI
Monthly PrEP Use	Self-reported percent of time in the past month that the patient took PrEP as prescribed	CASI
Incident Sexually Transmitted Infections (STI): gonorrhea, chlamydia, and syphilis	Self-reported diagnosis of the STI in the past 3 months	CASI
Persistence	Reporting currently taking PrEP	CASI
Retention	Patient attended at least one routine/scheduled PrEP visit in the past 3 months	CASI
Sexual Risk	Self-reported number of times receptive condomless anal sex in the past 3 months Self-reported number of times insertive condomless anal sex in the past 3 months	CASI CASI

4 Analysis Plan

For all of the above objectives, targeted minimum loss-based estimation (TMLE) will be used to estimate the mean counterfactual outcome for the population of interest, under each treatment, and at each time point (3 months and 6 months). For binary variables we will estimate the proportion or probability of the outcome.

4.1 Notation and Causal Estimands

The following notation will be used to denote study data:

- A: randomized treatment assignment at baseline
- Y_j: outcome of interest at the jth time point
- W_0: patient baseline covariates
- Δ_{Yj} whether or not Y_j was measured for the patient (1 = observed, 0 = missing)

The subscript j will take on values 0, 1, 2, to denote data collected at baseline, 3 months and 6 months, respectively. We assume that the observed data may be represented sequentially for the three month outcome as $O_1 = (W_0, A, \Delta_{Y,1}, \Delta Y_1)$ and for the six month outcome as $O_2 = (W_0, A, \Delta_{Y,2}, \Delta Y_2)$.

In order to define our causal estimands of interest, we use counterfactual notation. Counterfactual notation describes the outcome a participant would have under a specified intervention (even if that intervention was not observed for the patient). We denote this outcome as $Y(a)$ to represent the outcome under treatment $A = a$. Our estimands of interest generally take the form $E[Y(1) - Y(0)]$ which is the expected difference in the average outcome if everyone in the population took treatment one versus treatment zero. We use the following notation to describe the assigned treatments: (a_1) control, (a_2) P3 and P3+, (a_{2a}) P3, and (a_{2b}) P3+. Table 6 outlines the estimands of interest for each study objective. Note that since a patient

Table 6: Estimands of Interest.

Study Objective	Estimand[a]	Comparison	Outcome(s)
Primary	$P[Y(a_2)] - P[Y(a_1)]$	Average effect of P3/P3+, compared to SOC	Protective levels of TFV-dp and FTC-tp in DBS
Secondary I	$P[Y(a_{2a})] - P[Y(a_1)]$ $P[Y(a_{2b})] - P[Y(a_1)]$	Average effect of P3 compared to SOC Average effect of P3+ compared to SOC	Same as primary
Secondary II	$P[Y(a_2)] - P[Y(a_1)]$	Average effect of P3/P3+ compared to SOC	Self-reported weekly PrEP use Self-reported monthly PrEP use
Secondary III	$P[Y(a_2)] - P[Y(a_1)]$	Average effect of P3/P3+ compared to SOC	Panel of self-reported outcomes
Cost	$\frac{E[C(a_{2b}) - C(a_{2a})]}{E[Y(a_{2b}) - Y(a_{2a})]}$	Ratio of expected cost difference to expected efficacy difference between P3+ and P3	C will be calculated cost and Y will be the outcome of interest for assessing treatment efficacy

[a] The difference in notation between P and E represents whether the estimand is a difference in proportions or a difference in mean outcomes, respectively.

cannot theoretically receive both P3 and P3+ as treatment, the counterfactual outcome under treatment (a_2) represents the average counterfactual outcome of patients under P3 and P3+.

In order to estimate the estimands in table 3, we may use our observed data and, under certain assumptions, obtain estimates via the g-formula (see Supplement for details) and TMLE.

4.2 TMLE Analysis for Average Causal Effects

TMLE, combined with super learning, will be used to address both the primary and secondary study objectives. The regression models needed for TMLE estimation include the propensity for treatment, the propensity for missingness, and the outcome regression. We will use super learning to estimate these regressions. Super learning is a machine learning method that, when given a set of modeling algorithms, uses cross-validation to determine the optimal convex combination of the algorithms. This allows us to combine flexible learners with more traditional methods (linear/logistic regression) to create our models, and thereby improves the chances that the models are specified correctly. The details for building each regression are given below.

Estimating Propensities

The probability of treatment assignment at baseline is essentially known since this is a randomized trial. Therefore, we will use simple empirical means to estimate the probability of treatment $A = a$. Using the notation $\bar{g}_a$ to denote the propensity for treatment a we may write.

$$\bar{g}_{a,n} = \frac{1}{n} \sum_{i=1}^{n} I(A = a)$$

Where the subscript n is added to indicate that this is an estimator of $\bar{g}_{a,0}$.

Unlike treatment assignment, the propensity of missing a follow-up time point is not randomized and may be estimated with parametric models or via data adaptive methods. From henceforth, we will denote the propensity of having data at follow-up time point j, as $\bar{g}_{\Delta_{Yj}}(a, w_0) = P(\Delta_{Yj} = 1 \mid W_0 = w_0, A = a)$, where j= 1,2 and a is the treatment regimen of interest. To create this model, we will regress Δ_{Yj} on baseline patient covariates and dummy variables for each possible treatment group. For baseline patient covariates, we will use age, baseline DBS data, baseline PrEP use (yes or no), site, and race/ethnicity. We will employ super learning with a candidate library to estimate the propensity for missingness. See table 7 for candidate algorithms and the variables considered.

Estimating Outcome Regressions

In order to estimate the mean counterfactual outcome via the g-formula, we must estimate an outcome regression. Suppose our objective has K treatments of interest then k = 1, 2, ..., K. Let k denote the k^{th} treatment of interest. The necessary regression for the outcome at time point j, j=1,2, takes the following form:

$$\bar{Q}_k^j(w_0) = E(Y_j \mid A = a_k, \Delta_{Yj} = 1, W_0)$$

Similar to our approach to propensity score estimation, we will estimate this regression with super learning and will include all treatment interventions A, by incorporating a dummy variable for each possible intervention (except control group which will be the reference group). Each regression will be fit on a subset of data where $\Delta_{Yj} = 1$. The baseline covariates included will be age, DBS, PrEP use, site, and race/ethnicity. See table 7 for the super learner libraries and variables included.

Once these regressions are fit, we use the TMLE estimator of average treatment effect (ATE) to obtain our final estimates for the counterfactual outcomes of interest. For additional details regarding the regressions fit, super learning, TMLE, and variance estimation please refer to the Supplement. For the primary objective please also refer to section 7.2 and Supplement A.5 for additional considerations regarding the TMLE analysis.

4.3 Hypothesis Tests and Multiple Testing Adjustment

Table 8 lists the hypotheses to be tested for each study objective. Each hypothesis comparing P3/P3+ to SOC will be tested at both 3 and 6 months, separately. For comparisons of P3 vs SOC and P3+ vs SOC, we will report the comparisons in a descriptive manner as opposed to conducting additional hypothesis tests (i.e. Secondary objectives I, II, and III).

Table 8 lists the p-value adjustment for each hypothesis test. For the primary objective no adjustment is done. For secondary objective II, a Bonferroni Correction will be applied to adjust for multiple testing and to maintain a nominal alpha level of approximately 0.05 within the objective. Due to the exploratory nature of secondary objective III, we will use a Benjamini-Hochberg adjustment, which controls the false discovery rate (FDR). We will use 0.20 as the desired FDR, which is the proportion of rejected null hypotheses that are actually true (See Glickman, 2014 for more information).

4.4 Cost Analysis

Cost was reported at the site-level. To calculate the incremental cost effectiveness ratio we assigned to each participant the average cost per person at their study site and within their treatment arm. We treated this value as if it was the participant's realized cost. This was then used in a complete case analysis comparing the mean cost difference between participants on P3+ vs P3 divided by the mean efficacy difference as measured by the binary TFV measure at 3 months. To create a 95% confidence interval we used the percentile bootstrap method.

Table 7: Super Learner Libraries and variables used for the propensity score and outcome regressions

Model	Outcome	Learner(s)	Baseline Variables[1]
$\bar{g}_{\Delta_{Y^j}}(a, w_0)$	Self-Reported PrEP use or Exploratory Outcome of Interest Observed	SL.glm	BEO[2], BEO*trt
			Age, BEO, On PrEP, Site, Race/Ethnicity
			Age, BEO, on PrEP, Site, Race/Ethnicity, (BEO * On PrEP), (Age*trt)
		SL.earth	Age, BEO, on PrEP, Site, Race/Ethnicity
		HAL	Age, BEO, on PrEP, Site, Race/Ethnicity
		glmnet	Age, BEO, on PrEP, Site, Race/Ethnicity, (BEO * On PrEP), (Age*trt)
		step	Age, BEO, on PrEP, Site, Race/Ethnicity, (BEO * On PrEP), (Age*trt)
$\bar{Q}_k^j(w_0)$	Self-Reported PrEP use or Exploratory Outcome	SL.glm	BEO, BEO*trt
			Age, BEO, On PrEP, Site, Race/Ethnicity
			Age, BEO, on PrEP, Site, Race/Ethnicity, (BEO * On PrEP), (Age*trt)
		SL.earth	Age, BEO, on PrEP, Site, Race/Ethnicity
		HAL	Age, BEO, on PrEP, Site, Race/Ethnicity
		glmnet	Age, BEO, on PrEP, Site, Race/Ethnicity, (BEO * On PrEP), (Age*trt)
		step	Age, BEO, on PrEP, Site, Race/Ethnicity, (BEO * On PrEP), (Age*trt)

trt - Treatment
BEO - Exploratory measure at baseline
[1] Baseline variables to include in the model, in addition to dummy variables for treatment
[2] If the outcome of interest is available at baseline, it will be controlled for. If not, another variable may be controlled for, if it is predictive for both the EO and missingness.

Table 8: Hypotheses and P-Value Significance Thresholds

Study Objective	Hypothesis	Significance Threshold
Primary	There is no difference between P3/P3+ and SOC in terms of PrEP adherence, as measured by binary FTC-tp. There is no difference between P3/P3+ and SOC in terms of PrEP adherence, as measured by binary TFV-dp.	0.05
Secondary II	There is no difference between P3/P3+ in terms of self-reported weekly PrEP use. There is no difference between P3/P3+ in terms of self-reported monthly PrEP use.	0.025
Secondary III	There is no difference between P3/P3+ and SOC in terms of self-reported outcome.	Benjamini-Hochberg

Table 9: Baseline variables stratified by treatment group (Table 1)

Variable	P3 (N=)	P3+ (N=)	SOC (N=)
Symmetric Baseline Variables, mean (SD)			
Skewed Baseline Variables, median (Q1,Q3)			
Categorical Baseline Variables, count (%)			

5 Reporting Results

5.1 Initial Descriptive Statistics and Data Exploration

Before conducting a formal causal analysis using the g-formula and assumptions described above, we will first use descriptive statistics to describe the study cohort and understand our study variables. Table 1, will include patient baseline characteristics (including demographic, social, baseline sexual risk variables, PrEP use, and study site) for the entire cohort, stratified by treatment group (see table 9 for a shell table).

To visualize differences in study outcomes by treatment group for both the primary objective and secondary objectives I and II, we will plot line graphs, with crude average trend lines for each treatment group. For continuous outcomes individual trend lines may also be plotted on these graphs.

5.2 Primary and Secondary Analysis Results

The results of these analyses will be presented in four separate tables. The tables will be organized as follows:

- Table 2 will contain results pertaining to the primary objective, see shell table 10.
- Table 3 will contain results pertaining to the secondary objective I, see shell table 11.
- Table 4 will contain results pertaining to the secondary objective II, see shell table 12.
- Table 5 will contain results pertaining to the secondary objective III, see shell table 13. For any significant results, we may also create tables to present comparisons of P3 vs SOC and P3+ vs SOC separately.

Table 10: Estimated difference in proportion of protective PrEP use between treatment group and SOC group as measured by FTC-tp and TFV-dp. (Table 2)

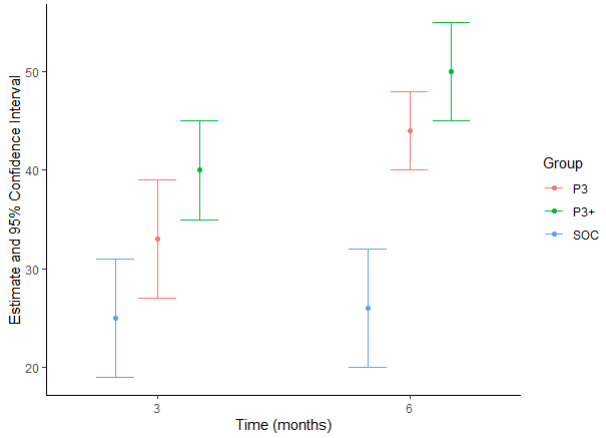
P3/P3+ vs. SOC	Estimated Difference	95% CI	p-value
Protective PrEP Level in Blood: TFV-dp			
3 month			
6 month			
Protective PrEP Level in Blood: FTC-tp			
3 month			
6 month			

- Table 6 will contain results pertaining to the cost effectiveness objective.

As mentioned previously, if a formal hypothesis test was not used to compare groups, we will report results descriptively.

In addition to tables we will also visualize results by plotting point estimates and 95% confidence intervals. The plot will contain the average counterfactual outcome on the y-axis, and 3 and 6 months on the x-axis. Each treatment group will have a separate point on the plot. See figure 1 for an example plot.

Figure 1: Example plot for displaying estimates.



6 General Considerations and Statistical Issues

6.1 Timeline

1. Data Available: February 24, 2022
2. Analysis Done: April 20, 2022
3. Manuscript Complete: TBD

Table 11: Estimated difference in proportion of protective PrEP use between treatment groups and SOC group, where P3 and P3+ are considered separately (table 3)

P3 vs. SOC	Estimated Difference	95% CI
Protective PrEP Level in Blood: TFV-dp		
3 months		
6 months		
P3+ vs. SOC	Estimated Difference	95% CI
Protective PrEP Level in Blood: TFV-dp		
3 months		
6 months		
P3 vs. SOC	Estimated Difference	95% CI
Protective PrEP Level in Blood: FTC-tp		
3 months		
6 months		
P3+ vs. SOC	Estimated Difference	95% CI
Protective PrEP Level in Blood: FTC-tp		
3 months		
6 months		

Table 12: Estimated difference in self-reported PrEP use between treatment groups and SOC group.

P3/P3+ vs SOC	Estimated Ratio	95% CI	p-value
Self-reported Weekly PrEP Use			
3 months			
6 months			
Self-reported Monthly PrEP Use			
3 months			
6 months			
P3 vs SOC	Estimated Ratio	95% CI	
Self-reported Weekly PrEP Use			
3 months			
6 months			
Self-reported Monthly PrEP Use			
3 months			
6 months			
P3+ vs SOC	Estimated Ratio	95% CI	
Self-reported Weekly PrEP Use			
3 months			
6 months			
Self-reported Monthly PrEP Use			
3 months			
6 months			

Table 13: Estimated difference in proportions (or mean outcomes) between treatment group and SOC group for self-reported exploratory outcomes (table 5)

P3/P3+ vs SOC	Estimated Difference	95% CI	p-value
Outcome 1			
3 months			
6 months			
Outcome 2			
3 months			
6 months			

Table 14: Frequency of Missing Data, in entire cohort and by treatment group, reported as count (%). (table 7)

Variable	Total (N=)	P3 (N=)	P3+ (N=)	SOC (N=)
Baseline data				
Baseline Variables				
3 month data				
TFV-dp				
FTC-tp				
Outcome 1				
...				
Outcome p				
6 month data				
TFV-dp				
FTC-tp				
Outcome 1				
...				
Outcome p				

6.2 Missing Data and Modification of Primary Analysis

Due to COVID-19, the collection of dry blood spots (DBS) was interrupted during this study. Therefore, the missingness in DBS will be high and it is possible for patients to have follow-up survey data, while missing the DBS follow-up data, or vice versa. Although TMLE can accommodate missingness at random (conditioned on baseline covariates) in the outcome of interest, a high amount of missing data in DBS will increase the variance of the estimator. Therefore, we decided to use survey data to "impute" the outcome variable. In doing so, we also require additional regression models to build the estimator for ATE. Please refer to the Supplement for details.

There are two survey variables that will be used to impute the values of the DBS data, namely PrEP use in the past week (number of days) and PrEP use in the past month (reported as a percentage of days). If the patient reported that they are *not* currently on PrEP, then the weekly PrEP use is assumed to be zero.

It is possible for one survey question to be present, but the other to be missing. In these cases we will use three simple forms of imputation to deal with the missing survey variable:

1. We will use mean imputation. First we stratify by current PrEP use (Y/N) and then within each group we perform mean imputation for the missing variable.

2. We will predict the survey variable (weekly or monthly PrEP use) with the other survey variable that is present using simple linear regression models

3. To impute weekly PrEP use from monthly PrEP use, we scaled the participant's month variable to range from 0-7. Similarly to impute monthly PrEP use from weekly PrEP use, we scaled the participant's week variable to range from 0-100.

Once this imputation is performed we can build a super learner library that includes each of the imputation methods and allow the super learner to pick the"best" methods via giving them weight in the final learner.

In addition to missingness in DBS and survey data, there may be missingness in other baseline measures or follow-up data of interest. We will assess the data for missingness in other data points and report the frequency of missing data points (see shell table 14). The method for handling missing values will be determined once the severity of missingness is understood and after discussion with the research team to understand potential causes for the missing data. Options for handling this missing data include imputation or omitting the follow-up data for a patient (i.e. consider setting $\Delta_{Y,ij} = 0$ for the patient).

6.3 Confounder Selection

Since treatment was randomized in this study, our primary concern is controlling for confounders that affect both whether patients remain in the study and whether patients adhere to PrEP. W_0 and W_1 are potentially high-dimensional vectors, therefore, it is best advised to select a subset of the covariates to control for in the analyses. Factors identified by the study team that predict PrEP adherence are: age, race/ethnicity, PrEP stigma, substance use, mental health, prior PrEP use, PrEP self-efficacy, partner serostatus, and sexual behavior. In the interest of parsimony, we selected age, baseline DBS, site, race/ethnicity, and baseline PrEP use (yes/no) as the variables to control for in our analysis.

It is possible that there are time-varying confounders of the missingness/outcome relationship. Unfortunately, we cannot account for these confounders for our analysis of the three month time point. For the six month outcome, a confounder of the missingness/outcome relationship may be PrEP use at three months. One assumption of our primary analysis, is that this confounding does not exist or is negligible.

6.4 Definition of the Primary Outcome

To better understand the sensitivity of our primary analysis to how adherence to PrEP was measured, we also considered alternative definitions of adherence as a sensitivity analysis. The alternative definitions considered were *both* TFV-dp and FTC-tp levels in the blood consistent with PrEP use > 4 days per week and *at least one* blood marker, TFV-dp for FTC-tp, consistent with PrEP use > 4 days per week.

A Supplement

A.1 G-Formula for causal estimands of interest

The g-formula allows us to estimate our causal estimands with the observed data (under assumptions). For the counterfactual outcomes under treatment a_k, $E[Y_j(a_k)]$, where j=1,2, we may define the following quantities:

- $Q_1^j(y, w_0)$: the conditional CDF of Y_j given $A = a_k$, $\Delta_{Y,j} = 1$, and W_0
- $Q_0^j(w_0)$: the CDF of W_0.

With these CDFs, we may define the g-formula as:

$$E[Y_j(a_k)] = \int \bar{Q}_1^j(w_0) dQ_0^j(w_0)$$

where

$$\bar{Q}_1^j(w_0) = \int y dQ_1^j(y, w_0)$$

A.2 Causal Assumptions

In order to establish identifiability of the expected counterfactual outcomes via the g-formula written above, a few assumptions are necessary. We assume randomization which states that $A \perp\!\!\!\perp Y(a) \mid W_0$ and also that $\Delta_{Yj} \perp\!\!\!\perp Y(a) \mid W_0, A$. We know that the first randomization assumption holds because treatment was randomized in this study. For the second randomization assumption to be true, we assume that we have controlled for all confounders of missingness and outcome relationship via the baseline variables controlled for in our analyses (age, On PreP, and outcome of interest measured at baseline). This assumption also implies that there is no time varying confounding, or variables after baseline which affect the relationship between missingness and the outcome.

The second major assumption is the positivity assumption, which states that $P(A = a \mid W_0) > 0$ for all a and $P(\Delta_{Yj} = 1 \mid W_0, A) > 0$. This assumption states that each participant has a non-zero probability of receiving treatment a (which is true since this is a randomized trial), and a non-zero probability of having data for each follow-up time point, given treatment and baseline covariates. The last major assumption is the consistency assumption which states that for a patient whose treatment $A = a$, the outcome we observe is equal to the counterfactual outcome under treatment regimen a for that patient.

A.3 Super Learning and TMLE Parameters

Certain parameters should be specified a priori for using the super learner and TMLE. For both the outcome regression and the propensity for missingness we will use 10-fold cross validation. The "method" used will be CC_nloglik for binary outcomes and CC_LS for continuous outcomes. These methods minimize the negative binomial log likelihood on the logistic scale and the squared error loss, for binary and continuous outcomes respectively. The appropriate "family" argument will also be specified according to the distribution of the outcome.

A.4 Variance Estimation

We will obtain the influence curve for our estimators of interest, evaluated at each observation. This will allow us estimate the variance.

For example, let $IC_{j1}(O_i)$ and $IC_{j2}(O_i)$ denote the influence curve evaluated on observation i, for our estimator of $E[Y_j(a_{k1})]$ and $E[Y_j(a_{k2})]$, respectively. We know that the distribution for our estimators of $E[Y_j(a_{k2}) - Y_j(a_{k1})]$ under, the null hypothesis of no difference between treatment groups, will be normal with mean zero and variance σ_j^2, where:

$$\sigma_j^2 = Var((IC_{j1}(O_i) - IC_{j2}(O_i))$$

which may be estimated by:

$$\hat{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n}(IC_{j1}(O_i) - IC_{j2}(O_i))^2$$

A.5 Incorporating Survey Data into the Primary Analysis

As noted in the body of the text, self-reported PrEP use was incorporated into the primary analysis to inform our estimates due to a high level of missing dbs data. The details of this analysis are organized as follows: (1) notation, (2) the new estimand, (3) the efficient influence curve, (4) implementation steps, and (5) additional causal assumptions.

A.5.1 Notation

We define additional notation to use in this portion of the supplement. For variables, the time subscript j is dropped for convenience. The additional variables are defined as follows:

- Y_T: denotes the true underlying value of the outcome of interest ($Y = Y_T$ when $\Delta_Y = 1$)
- S: denotes observed survey variables, PrEP use in the past week and PrEP use in the past month observed at the same time as the outcome of interest.
- Δ_S: is an indicator variable for whether survey data is observed (1 if either self-report measure is available, 0 else)
- S_T: denotes the true survey variable ($S = S_T$ when $\Delta_S = 1$)

A.5.2 New Estimand

It can be shown, under certain independence and consistency assumptions, that:

$$\begin{aligned} E[Y(a)] =& E\Big[E[Y \mid \Delta Y = 1, W, A = a](1 - I(\Delta Y = 0, \Delta S = 1)) \\ &+ E\big[E[Y \mid \Delta Y = 1, \Delta S = 1, A = a, W, S] \mid \Delta S = 1, WA = a]I(\Delta Y = 0, \Delta S = 1]\Big] \end{aligned}$$

Therefore, we denote our estimand of interest as:

$$\Psi(P) = E\Big[E[Y \mid \Delta Y = 1, W, A = a](1 - I(\Delta Y = 0, \Delta S = 1)) + E\Big[E[Y \mid \Delta Y = 1, \Delta S = 1, A = a, W, S] \mid \Delta S = 1, WA = a]I(\Delta Y = 0, \Delta S = 1]\Big]\Big]$$

A.5.3 Efficient Influence Curve of the Estimand

We can derive the efficient influence curve for the right hand side of the above equation, which gives:

$$(y - \bar{Q}_a(w))\frac{I(A = a, \Delta Y = 1)}{P(A = a, \Delta Y = 1, \mid W = w)}\big(1 - P(\Delta Y = 0, \Delta S = 1 = w)\big) +$$
$$(y - \bar{Q}_a(s, w))\frac{I(A = a, \Delta Y = 1, \Delta S = 1)}{P(A = a, \Delta Y = 1, \Delta S = 1 \mid W = w)}P(\Delta Y = 0, \Delta S = 1 \mid W = w) +$$
$$(\bar{Q}_a(s, w) - E[\bar{Q}_a(s, w) \mid A = a, W = w, \Delta S = 1])\frac{I(A = a, \Delta S = 1)}{P(A = a, \Delta S = 1 \mid W)}P(\Delta Y = 0, \Delta S = 1 \mid W = w) -$$
$$\Psi(P) + \bar{Q}_a(W)(1 - I(\Delta S = 1, \Delta Y = 0)) + E[\bar{Q}_a(S, W) \mid A = a, \Delta S = 1, W]I(\Delta Y = 0, \Delta S = 1)$$

where $\bar{Q}_a(W) = E[Y \mid A = a, \Delta Y = 1, W]$ and $\bar{Q}_a(S, W) = E[Y \mid A = a, \Delta Y = 1, \Delta S = 1, S, W]$. The first three lines of the above equation can be targeted with TMLE.

A.5.4 Implementation Steps

The below steps are taken to build the TMLE estimator for ATE based on the estimand and its efficient influence curve. Table 15 gives the super learner algorithms and variables included for each model in the below steps.

1. Estimate the outcome regression with survey data added as a predictor, $\bar{Q}_a(S, W) = E[Y \mid S, W, A = a, \Delta_S = 1, \Delta_Y = 1]$ to get $\bar{Q}_{a,n}(S, W)$.

2. Estimate the propensity for observing the survey data, $\bar{g}_{\Delta_S}(A, W, \Delta_Y) = P(\Delta_S = 1 \mid A, W, \Delta_Y)$ to get $\bar{g}_{\Delta_S,n}(A, W, \Delta_Y)$. Also estimate the propensity for observing the outcome, $P(\Delta_Y = 1 \mid A, W)$, and the propensity for treatment, as previously described. These three propensities may be used to calculate the probabilities needed for steps 3 and 7 below, where $\tilde{P}$ is used to denote the estimates of the true probability distribution.

3. For each treatment of interest, $A = a$, perform a TMLE-style fluctuation of $\bar{Q}_{a,n}(S, W)$ to obtain $\bar{Q}^*_{a,n}(S, W, \epsilon_0)$. This may be done by fitting a logistic regression model with Y as the outcome and an offset term $logit(\bar{Q}_{a,n}(S, W))$ plus a clever covariate $\Delta_S\Delta_Y I(A = a)\frac{\tilde{P}(\Delta_S=1,\Delta_Y=0|W)}{\tilde{P}(\Delta_Y=1,\Delta_S=1,A=a|W)}$ for the mean model. ϵ_0 is used to denote the coefficient for the clever covariate.

4. Let's denote the Maximum Likelihood estimate of ϵ_0 from step 3 as $\hat{\epsilon}_0$. We may evaluate $\bar{Q}^*_{a,n}(S, W, \hat{\epsilon}_0)$ for those with $\Delta_S = 1$ and under $\Delta_Y = 1$ with the following equation:

$$Y^* = \bar{Q}^*_{a,n}(S, W, \hat{\epsilon}_0) = expit[(logit(\bar{Q}_{a,n}(S, W))) + \hat{\epsilon}_0 I(A = a)\frac{\tilde{P}(\Delta_S = 1, \Delta_Y = 0 \mid W)}{\tilde{P}(\Delta_Y = 1, \Delta_S = 1, A = a \mid W)}]$$

5. Define $\tilde{Y}$ as Y^* when $\Delta_Y = 0$ and $\Delta_S = 1$ and Y otherwise.

6. Estimate the model $\bar{Q}_a(W, \delta) = E[\tilde{Y} \mid W, A, \delta > 0, \delta]$ to get $\bar{Q}_{a,n}(W, \delta)$. where:

$$\delta = \begin{cases} 1 & \text{if } \Delta_Y = 1 \\ 0 & \text{if } \Delta_S = 1 \text{ and } \Delta_Y = 0 \\ -1 & \text{else} \end{cases}$$

7. Fluctuate $\bar{Q}_{a,n}(W,\delta)$ to get $\bar{Q}^*_{a,n}(W,\delta,\epsilon)$ by fitting two logistic regression models for each possible treatment $A = a$. Both logistic regression models will use $\tilde{Y}$ as the outcome, but different mean models (a) and (b) below.

   (a) $logit(\bar{Q}_{a,n}(W,\delta=1)) + \epsilon_1 \Delta_Y I(A=a)\frac{1-\tilde{P}(\Delta_S=1,\Delta_Y=0|W)}{\tilde{P}(\Delta_Y=1,A=a|W)}$

   (b) $logit(\bar{Q}_{a,n}(W,\delta=0)) + \epsilon_2 \Delta_S I(A=a)\frac{\tilde{P}(\Delta_S=1,\Delta_Y=0|W)}{\tilde{P}(\Delta_S=1,A=a|W)}$

8. For those with $\Delta_S = 1$ and $\Delta_Y = 0$, model (b) defines $\bar{Q}^*_{a,n}(W,\delta,\epsilon)$. Estimate $\bar{Q}^*_{a,n}(W,\delta,\epsilon)$ with fluctuation (b) for these individuals under $\Delta_S = 1$ and $A = a$ for the participant.

9. For everyone not included in step 8 fluctation (a) defines $\bar{Q}^*_{a,n}(W,\delta,\epsilon)$. Estimate $\bar{Q}^*_{a,n}(W,\delta,\epsilon)$ with fluctuation (a) setting $\Delta_Y = 1$ and $A = a$ for the participant.

10. Take the sample mean of $\bar{Q}^*_{a,n}(W,\delta,\epsilon)$ to estimate $E[Y(a)]$.

11. Calculate the treatment effect of interest in the analysis, as detailed in table 6.

12. The variance of the estimator for $E[Y(a)]$ may be estimated with the sample variance of the efficient influence curve.

A.5.5 Additional Causal Assumptions

There are additional assumptions that must be made in order to infer causality for the primary analysis, in addition to the aforementioned causal assumptions (Section A.2). (1) It is assumed that $P(\Delta_S = 1 \mid \Delta_Y = 1, W = w, A = a) > 0$ for each possible combination of $A = a$ and $W = w$, and (2) it is assumed that the survey data is missing at random. The specific independence assumptions necessary are listed in table 16. For these assumptions to be true, all confounders of missingness and the outcome relationship must be controlled for in the analysis.

B References

Benkeser D, Carone M, Van der Laan M, Gilbert P. Doubly robust nonparametric inference on the average treatment effect. Biometrika, 2017; 104(4): 863-880.
Glickman, Mark, Rao S, Schultz M. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. Journal of Clinical Epidemiology. 2014; 67: 850-857.

Table 15: Super Learner Libraries used in final analyses

Model	Outcome	Learner(s)	Baseline Variables[1]
$\bar{g}_{\Delta_{Y}j}(a, w_0)$	DBS data Observed	SL.glm	DBS, DBS*trt
			Age, DBS, On PrEP, Site, Race/Ethnicity
			Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt)
		SL.earth	Age, DBS, on PrEP, Site, Race/Ethnicity
		HAL	Age, DBS, on PrEP, Site, Race/Ethnicity
		glmnet	Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt)
		step	Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt)
$\bar{g}_{\Delta_{S}j}(a, w_0, \delta_Y)$	Self-reported PrEP use Observed	SL.glm	DBS, DBS*trt, Δ_Y
			Age, DBS, On PrEP, Site, Race/Ethnicity, Δ_Y
			Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt), Δ_Y
		SL.earth	Age, DBS, on PrEP, Site, Race/Ethnicity, Δ_Y
		HAL	Age, DBS, on PrEP, Site, Race/Ethnicity, Δ_Y
		glmnet	Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt)
		step	Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt)
$Q_a(s, w_0)$	PrEP DBS	SL.glm	DBS, DBS*trt,S
			Age, DBS, On PrEP, Site, Race/Ethnicity, S
			Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt), S
		SL.earth	Age, DBS, on PrEP, Site, Race/Ethnicity, S
		HAL	Age, DBS, on PrEP, Site, Race/Ethnicity, S
		glmnet	Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt)
		step	Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt)
$\bar{Q}_a(w_0, \delta)$	$\check{Y}$	SL.glm	DBS, DBS*trt
			Age, DBS, On PrEP, Site, Race/Ethnicity, δ
			Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt), δ
		SL.earth	Age, DBS, on PrEP, Site, Race/Ethnicity, δ
		HAL	Age, DBS, on PrEP, Site, Race/Ethnicity, δ
		glmnet	Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt), δ
		step	Age, DBS, on PrEP, Site, Race/Ethnicity, (DBS * On PrEP), (Age*trt), δ

Treatment
DBS - dry blood spot
[1] Baseline variables to include in the model, in addition to dummy variables for treatment

Table 16: Independence Assumptions.

Scenario	Analyses	Assumptions
Survey data incorporated	Primary and Secondary	$\Delta_Y \perp\!\!\!\perp Y_T \mid A, W$ $\Delta_S \perp\!\!\!\perp Y_T \mid A, W, \pm\Delta_Y$ $\Delta_Y \perp\!\!\!\perp Y_T \mid A, W, S, \Delta_S$ $\Delta_S \perp\!\!\!\perp S_T \mid A, W$ $\Delta_Y \perp\!\!\!\perp S_T \mid A, W, \Delta_S$
Survey data not incorporated	Exploratory	$\Delta_Y \perp\!\!\!\perp Y_T \mid A, W$