This is IMPAACT 2016 SAP Version 2.0 with names of authors, names of publication writing team members and analysis timeline redacted.

IMPAACT 2016

Primary Statistical Analysis Plan

Version 2.0

Evaluating a Group-Based Intervention to Improve Mental Health and ART Adherence Among Youth Living with HIV in Low Resource Settings

ClinicalTrials.gov Identifier: NCT04024488

Protocol Version 2.0

April 28, 2025

Table of Contents

Version History Table

Version	Changes Made	Date Finalized
1.0	Original Version	4 October 2019
1.1	Reviewed for protocol version 2.0. Added additional statistician; no other changes required. Study has not yet begun enrolling participants.	27 April 2023
2.0	Updated SAP to address protocol deviations and handling of missing or unscorable data. Added additional statistician.	28 April 2025

1 Introduction

1.1 Purpose

This Primary Statistical Analysis Plan (SAP) describes the primary and secondary outcome measures of the Randomized Trial (RT) component of the IMPAACT 2016 study that will be included in the primary analysis report and primary manuscript, and which address the major primary and secondary objectives of the study. Note: these objectives are all for time points at or before the 6-month study visit, which is the primary completion date (PCD).

The Primary SAP outlines the general statistical approaches that will be used in the analysis of the study. It has been developed to facilitate discussion of the statistical analysis components among the study team, and to provide agreement between the study team and statisticians regarding the statistical analyses to be performed and presented in the primary analysis report, including the results for the primary and secondary outcome measures that will be posted on ClinicalTrials.gov.

Detailed outlines of tables, figures, and coding descriptions that will be included in the Primary Analysis Report are included in the Analysis Implementation Plan (AIP).

Analyses for the Primary Analysis Report will be finalized once the last participant has completed the 6-month study visit, all queries have been resolved, and the study database closure/data lock has been completed.

Outlines of analyses for other objectives and outcome measures not included in the Primary SAP will be provided in a separate SAP for Other Outcome Measures.

1.2 Version History

Version 2.0 of the Primary SAP provides additional detail on analytic strategies to address protocol deviations involving eligiblity, enrollment, or randomization and missing or unscorable data based on SMC request. A summary of major changes is as follows:

(1) Defines analysis populations and specifies their use in primary, secondary, and supplementary analyses;
(2) Provides detail on existing individual-level supplementary analyses, and adds new supplementary analysis for two-component, rather than three-component, mental health composite measure;
(3) Adds new supplementary analyses to investigate bias due to missing or unscorable data and eligibility, enrollment, or randomization deviations;
(4) Defines unscorable UCLA PTSD-RI assessments and specifies their handling in analysis, including which analyses will follow complete case and imputation strategies;
(5) Updates reporting to include number and nature of eligibility, enrollment, and randomization errors and reflect updated data collection.

2 Study Overview

2.1 Overview of Study Design

IMPAACT 2016 is a multi-site, two-arm, individually randomized group trial (IRGT) preceded by Focus Groups and Pilot Testing to adapt the intervention to the local context. The purpose of the study is to evaluate whether an Indigenous Leader Outreach Model (ILOM) of trauma-informed cognitive behavioral therapy (TI-CBT) is associated with improved mental health outcomes and antiretroviral therapy (ART) adherence among youth living with HIV. The study population consists of 15-19 year old youth living with HIV and mental health distress, along with their caregivers (if available and agreed to by the associated youth participant).

There will be approximately 192-256 youth participants (96-128 per arm), plus their caregivers, in the Randomized Trial. Youth participants will be randomized in a 1:1 ratio to one of two study arms (TI-CBT Intervention or Discussion Control) in a parallel-group design. Randomization will be stratified by gender to assure approximate balance between study arms. Each arm will have 12-16 groups with an average of eight youth participants per group (range 6-10). Youth participants' caregivers will be assigned to the same arm as their youth participants. For accrual, as a site recruits enough youth to form two groups, the youth will be randomized on the day of enrollment to either TI-CBT or Discussion Control.

Prior to the Randomized Trial, Focus Groups comprised of 5-8 youth participants and 5-8 caregiver participants will be conducted at selected sites. All sites will also conduct a Pilot Test with up to eight youth participants and up to eight caregiver participants. There will be no data analysis of the Focus Groups or Pilot Tests, which are therefore not included in this SAP.

Youth participants in the Randomized Trial will have the following study visits: screening (determine eligibility), pre-entry (collect baseline assessments), entry (randomization and enrollment), immediately post-last group session (IPL), 6-month follow-up, 12-month follow-up/exit visit. Caregiver participants will have the following study visits: screening/entry (may occur same day), IPL, 6-month follow-up, 12-month follow-up/exit visit. CASI-administered assessments will occur for all participants at the other visits; youth participants will also provide blood and hair samples. Youth participants will have six group sessions after entry while caregivers will have two. All participants will have a 6-month booster group session after the 6-month follow-up assessment.

2.2 Hypothesis

The TI-CBT Intervention will result in reduced symptoms of depression, anxiety, and/or traumatic stress compared to the Discussion Control at six months.

2.3 Study Objectives

This Primary SAP addresses the following primary and secondary objectives listed in the study protocol which will be included within the primary manuscript. Other study objectives in the protocol will be addressed in subsequent analysis plans.

The study objectives below will be analyzed under a superiority framework. As the primary completion date at 6 months precedes the end of follow-up at 12 months, the following objectives will only be considered prior to, or at, 6 months.

2.3.1 Primary Objective

1. Evaluate whether a TI-CBT Intervention is associated with improved depression, anxiety, and/or traumatic stress symptoms for youth living with HIV compared to a Discussion Control at six months [Objective 2.1.1].

2.3.2 Secondary Objectives

1. Evaluate whether a TI-CBT Intervention is associated with improved depression, anxiety, and/or traumatic stress symptoms for youth living with HIV compared to a Discussion Control following the last group session (at IPL) and at 12 months [Objective 2.2.2].
2. Assess whether a TI-CBT Intervention is associated with improved ART adherence (hair samples, self-report) and viral suppression (HIV RNA plasma) for youth living with HIV compared to a Discussion Control following the last group session (at IPL) and at 6 and 12 months [Objective 2.2.3].
3. Compare the rates of all targeted adverse events between a TI-CBT Intervention and a Discussion Control for youth living with HIV [Objective 2.2.8].

2.4 Overview of Sample Size Considerations

The proposed sample size for the Randomized Trial is 192-256 youth, expected to be enrolled from eight sites in four countries in sub-Saharan Africa. The sample size was chosen to provide 80% power to detect an effect size of 0.42-0.62 standard deviations between arms at 6 months, assuming an average group size of 8 and an intracluster correlation (ICC) range of 0.05 to 0.15.

This is a superiority study and the sample size was chosen to provide 80% power with a two-sided α=0.05 to detect a clinically important difference in the primary efficacy measures between study arms, allowing for a 10% loss to follow-up between study visits. The team will use two-sided tests to ensure power to detect differences between groups in both directions, but the hypothesis is that TI-CBT groups will lead to better mental health that the Discussion Control groups.

Further details on the assumptions and sample size calculations are provided in protocol Section 9.4.

2.5 Overview of Formal Interim Monitoring

An independent IMPAACT Study Monitoring Committee (SMC) will review this study regularly, following policies described in the IMPAACT Manual of Procedures. Once enrollment has started, there will be an SMC review to monitor accrual 6 months after the first participant is enrolled in the Pilot Study. Full SMC reviews will occur annually after the start of enrollment. The intervening 6-month reviews will focus on study milestones, accrual, fidelity to study interventions, and a qualitative discussion of feasibility. *Ad hoc* reviews will be scheduled as needed if any safety issues or concerns arise.

There will be no interim analysis of study outcome measures. However, there will be interim analyses to check the validity of the assumptions used for sample size calculations, to determine if it is necessary to increase the number of participants to ensure adequate power for the primary study objectives. Power decreases as the intra-cluster correlation (ICC, measure of correlation between treatment group members) increases. An interim evaluation of the ICC will occur once five groups in each Randomized Trial study arm have completed their six sessions, to see if it is greater than 0.15 (the largest value for which the selected sample size provides 80% power to detect acceptable minimum differences). An evaluation of loss to follow-up will occur when five groups/arm have reached the end of the window for their 6-month follow-ups. Both of these analyses will be presented to the SMC only, in a Closed Report. When five groups/arm have reached the end of the window for their 6-month follow-ups, the proportion of enrolled participants in the Randomized Trial who do not have six-month follow-up assessments will be assessed to see if it exceeds 10%. At each of these milestones, the protocol statistician will provide an interim analysis report including relevant data and assessment outcome, with a recommendation to maintain or increase the sample size for the SMC to review and make a final determination. If the SMC determines that the ICC or loss to follow-up rate require a sample size recalculation, the team will work with a blinded non-protocol statistician to determine the new sample size.

3 Outcome Measures

The study primary and secondary outcome measures used to address the study's primary and major secondary objectives are listed in Protocol Section 9.2 Table 4.

3.1 Primary Outcome Measures

The primary outcome measures used to address the study's primary objective are the following measures of mental health at 6 months after entry:

1. Patient Health Questionnaire-9 score (PHQ-9, measure of depression);
2. General Anxiety Disorder-7 score (GAD-7, measure of anxiety);
3. UCLA Post-Traumatic Stress Disorder-Reaction Index (UCLA PTSD-RI, measure of PTSD);
4. Composite mental health measure (PHQ-9, GAD-7, and UCLA PTSD-RI scores will be standardized and then summed).

3.2 Secondary Outcome Measures

The secondary outcome measures used to address the secondary objectives listed in Section 2.3 are as follows.

Mental Health Measures

The secondary mental health measures will be the same measures as in Section 3.1 above but assessed at the immediate post-last group session (IPL).

ART Adherence Measures

The following ART adherence measures will be assessed at the IPL and at 6 months:

1. Wilson 3-item scale (self-report);
2. Viral load (HIV-1 RNA copies/mL).

Note: Hair samples will be tested following the 12-month visit, and will not be available for the primary analysis report or primary manuscript.

Safety Measures

The outcome measures for safety are the following youth adverse events:

1. Grade 2 or higher suicidal ideation or attempts;
2. Grade 3 or higher psychiatric disorder;
3. Grade 3 or higher insomnia.

4 Statistical Principles

4.1 General Considerations

Baseline characteristics will be summarized by arm, but there will be no statistical comparisons comparing arms because of the randomized study design. There are no planned interim efficacy analyses.

All statistical tests will be two-sided and will not be adjusted for multiple comparisons. Categorical data will be summarized using N (%) and continuous data using N, min, Q1, median, Q3, max, and mean (standard deviation (SD)) (when appropriate). Any modifications to outcome measures after the team has seen data that were collected after entry will be identified as such in the analysis report.

NIH requires the primary analyses of treatment comparisons to be summarized by sex and by race/ethnicity and that treatment interactions with sex/gender and race/ethnicity will be tested. It is expected that all participants will be Black Africans, therefore treatment interactions by race/ethnicity will not be performed unless that proves incorrect. Note, however, that the study has not been specifically designed to provide high power to evaluate treatment differences within subgroups or to evaluate treatment by subgroup interactions.

4.1.1 Analysis Populations

We define two analysis populations:

- Intention-to-treat (ITT): All randomized youth, regardless of receipt of their randomized intervention.
- Modified Intention-to-Treat (MITT): Randomized youth who were not enrolled in error, regardless of receipt of their randomized intervention.

All primary, secondary, and supplementary analyses will be conducted using the ITT analysis population, as indicated by Protocol Section 9.5. Supplementary analyses may additionally use the MITT analysis population, with greater detail and justification provided in Section 4.2.1.

4.1.2 Times Used in Primary and Secondary Outcome Measure Definitions

Baseline (pre-entry visit): The day baseline evaluations are completed (0-14 days prior to the Entry Visit).

Day 0 (entry visit): Day of enrollment, randomization, and Group Session 1.

Immediately post-last group session (IPL): Day of follow-up evaluation (0-30 days after Group Session 6).

6-month follow-up visit: Day 182 ± 30 days.

12-month follow-up visit: Day 365 ± 30 days (not included in this SAP).

4.2 Analysis Approaches

4.2.1 Analysis of the Primary Objective

The primary efficacy analysis will compare all primary outcome measures between randomized arms using the ITT analysis population. The primary outcome measures will be analyzed using a two-sided type I error rate of 0.05 and associated 95% confidence intervals in comparing differences between randomized arms. There will be no statistical adjustment for multiple comparisons. Only complete cases, i.e., youth with data for a specified measure, will be analyzed.

It is expected that the three mental health measures (PHQ-9, GAD-7, UCLA PTSD-RI) will be correlated; they will therefore be analyzed jointly using a composite measure. The three measures will first be standardized to the same scale, with mean 0 and standard deviation 1. Standardization will be done for each measure by subtracting the baseline (Pre-Entry) overall variable mean from each observation and dividing by the baseline pooled sample standard deviation. Resulting standardized scores thus represent the number of standard deviations from the baseline mean. Internal consistency of the mental health composite will be checked by examining Cronbach's alpha for the three standardized mental health measures to see if they form a coherent measure, indicated by an alpha greater than or equal to 0.70. If coherent, the composite mental health score will be created by summing the standardized scores within-youth.

The primary analysis comparing the composite mental health measure between arms will be a group-level analysis. One valid method for analyzing intervention effects within cluster randomized trials is to compute the mean for each cluster and then perform a t-test on the cluster-specific summaries. Here, a paired t-test will be performed for the difference in means between arms, with pairs defined by each distinct combination of site and wave. The same procedure will be used to analyze each mental health measure separately.

Supplementary Analyses

Group treatment results in correlated responses due to participant interactions with a shared group of facilitators and participants. Failure to account for this correlation can increase false positive (type I error) rates. Furthermore, each site may feature up to two waves (Figure 1). Within each wave, one group of participants receives experimental TI-CBT while the other receives Discussion Control. Waves are subsequent, rather than concurrent, and group leaders are common between waves and within site. Thus, wave may reflect temporality, *e.g.*, treatment effects may change as leader experience increases, as well as between-group differences due to changes in group composition. Between-site heterogeneity is also expected. Multi-level models are commonly used to account for correlation within effect estimation, and also permit adjustment for wave and site, as well as additional covariates which can increase efficiency.

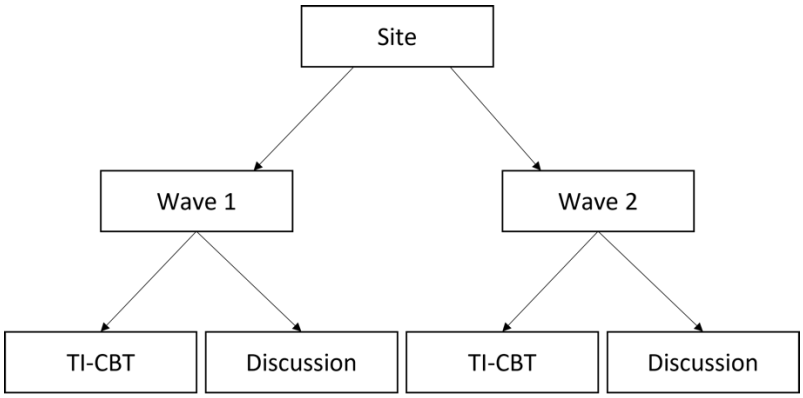

Figure 1. Structure of study site featuring two waves.

Supplementary analyses will be conducted separately for all primary outcome measures within the ITT analysis population using a multi-level model with two levels: individuals (level-1) nested within groups (level-2). Wave and site will be treated as fixed effects (i.e., adjusted for as covariates), along with other characteristics, as adjustment for strong prognostic factors can increase efficiency. Small-sample adjustment will be performed to ensure that inference is valid at the 5% significance level.

Potential covariates may include gender, number of sessions attended, caregiver participation, living situation (with or without parents), route of infection, and reported contamination between groups. Variable selection techniques will be used to identify candidate predictors strongly associated with each outcome among the ITT analysis population. Variables with $p<0.20$ within single variable models will be considered candidates for the final adjusted model. Candidate predictors will be selected into the final model following discussion with the protocol team to confirm clinical relevance (or prior evidence) and exclude (or combine) one or more covariates if highly correlated or colinear.

The GAD-7 and PHQ-9 assessments are scorable for all youth, unlike the UCLA PTSD-RI (see Section 4.2.3). If more than 10% of UCLA PTSD-RI assessments are unscorable at the PCD, another supplementary analysis will repeat the primary analysis for an additional composite

mental health measure compromised only of the GAD-7 and PHQ-9, regardless of Cronbach's alpha.

To investigate sources and impacts of bias, primary analyses will also be repeated (1) within the MITT analysis population, described below; and (2) using imputed data, described in Section 4.2.3, rather than a complete case approach.

ITT analyses are the only analyses protected by randomization, i.e., both measured and unmeasured confounders will be balanced between randomized arms thus providing an unbiased treatment effect, and represent the analysis standard for RCTs. However, in the presence of multiple eligibility or enrollment errors, the ITT analysis population may not reflect the target analysis population of IMPAACT 2016. For example, mental health symptoms may be less severe than expected if youth were enrolled with mental health scores less than the eligibility thresholds (i.e., PHQ-9 ≥ 10, GAD-7 ≥ 10, UCLA PTSD-RI ≥ 35). To assess the degree to which external generalizability of trial findings may be impacted by these protocol deviations, analysis will be repeated among the MITT analysis population which excludes them, and effects estimated among the ITT and MITT analysis populations will be compared. Hypothesized mechanisms for any observed differences between effects will be discussed in publications and presentations of trial results.

4.2.2 Analysis of the Secondary Objectives

Group-level analyses for the secondary mental health and adherence outcome measures will be the same as those used for the primary analysis of the primary outcome measures. The Wilson 3-item scale and log viral load are continuous measures. For viral suppression, which is a dichotomous measure, the first stage of the group-level analyses will involve calculating the proportion with that outcome within each group, and then proceeding as above using paired t-tests.

For the analysis of the secondary safety outcome measure, the number and proportion of youth experiencing at least one safety outcome measure (see Section 3.2) will be summarized by arm with exact 95% confidence intervals (CIs). Differences in proportions between arms will be summarized with an exact 95% CI on the difference. Proportions will also be summarized only including events assessed as related to treatment. The types of safety outcome measures that occur, including social harms, will be summarized.

4.2.3 Missing and unscorable data

<u>Loss-to-follow-up</u>

In mental health research, there is concern about biased results due to loss to follow-up. For that reason, if more than 10% of participants are lost to follow-up before six months, or if a larger proportion of participants are lost to follow-up in one arm (especially if the loss is due to serious adverse events, or suicide attempts), analyses will be undertaken to explore the potential effects of missing data on the conclusions of the study. The characteristics at study entry (gender, age, mode of transmission, severity of mental health distress and viral load suppression status) for participants who discontinued from the study before six months will be compared between the TI-

CBT and Discussion Control arms. Characteristics of participants discontinuing vs. not discontinuing will be compared overall and within intervention and control arms. In addition, the reasons for losses to follow-up will be assessed and compared between intervention and control arms. The probability and timing of loss to follow-up will be summarized and compared between arms using Kaplan-Meier plots and the log-rank test.

Unscorable UCLA PTSD-RI assessments

The UCLA PTSD-RI instrument used to assess trauma symptoms features two parts: (1) endorsement of one or more past traumas; and (2) assessment of post-traumatic stress symptoms, from which a symptom score is derived. If trauma is not endorsed in the first part, the second part is skipped and no score is obtained although the survey is technically "completed." We refer to such instances as "unscorable" assessments.

Unscorable assessments will be effectively treated as missing data. All primary and secondary objectives will be first conducted according to the complete case principle, wherein only youth with complete and scorable outcomes are included within analyses that include UCLA PTSD-RI scores. However, youth with unscorable assessments are likely to differ from those with scorable assessments, and these differences can bias treatment effect estimates. Restriction to complete cases also reduces sample size, and subsequently, power to detect a treatment effect. Imputation, or replacing unscorable data with a predicted score, can help with assessing the degree of bias within the complete case analysis and potentially improve power by allowing all youth to be included within analysis.

In what follows, missing refers to "unscorable." There are three types of missing data. If scores are "missing completely at random" (MCAR), observed scores can be thought of as a simple random sample of all scores. That is, missing scores are not related to any other missing or observed scores or variables. On the other hand, if scores are "missing not at random" (MNAR), whether a score is missing depends on the value of the score itself. Analyses of MNAR data are complex and require correct specification of the relationship between missing and observed scores, which is unknowable. An intermediate situation is if scores are "missing at random" (MAR), where missingness is assumed to depend only on observed information, such as other scores or participant and site characteristics. If the MAR assumption holds, and missing scores are predicted ("imputed") based on these observed factors, resulting analyses will be unbiased. Most analyses of missing data assume MAR.

MCAR is the only missingness assumption that can be assessed using the data. A missing indicator is constructed for each participant, equal to 1 if their survey was unscorable and 0 otherwise. If MCAR holds, all other variables should be approximately balanced between groups defined by the missing indicator. We will assess associations descriptively and formally using hypothesis testing within the ITT population only. Complete case analysis (CCA), which excludes youth with missing data, will be reported regardless of results. If investigation provides evidence that MCAR holds, the CCA will provide unbiased effect estimates. However, it is rare that data are truly MCAR.

If investigation provides evidence that the data is not MCAR, MAR will be assumed and imputation will be performed for the ITT analysis population. The set of variables used for

imputation will be informed by the analysis model and identified associations between missing indicators and observed variable, and finalized with input from the protocol team. Imputation will replace missing scores using predictions based on these variables. For example, UCLA PTSD-RI score may be predicted using PHQ-9 score, GAD-7 score, and site. To account for uncertainty in predicted values, multiple imputations will be performed. Analysis will be repeated for each imputed data set, and results averaged. Variance estimates will be obtained by applying "Rubin's Rules." Correlation among youth in the same group, or "clustering," will also be incorporated into imputation models.

5 Report Contents

5.1 CONSORT DIAGRAM

A flow diagram based on all screened youth which will include, but not be limited to, the following:

- Number of youth screened
- Number correctly found eligible
- Number erroneously found eligible by reason (e.g., PHQ-9 = 9)
- Number enrolled and randomized to each arm
- Number that did not receive randomized intervention
- Number on each arm with IPL data
- Number on each arm with 6-month data

5.2 Summaries

Detailed descriptions of the content of each of the following sections are given in the AIP.

5.2.1 Study entry

The following will be summarized by month and arm:

- Number of youth screened
    - Demographics of those found ineligible
    - Reasons for non-enrollment
- Number of youth enrolled

The following will be summarized by arm:

- Number of caregivers for whom youth participants approved participation
- Number of caregivers enrolled

There will also be a summary of eligibility and randomization errors overall.

5.2.2 Baseline characteristics

The following will be summarized by arm and overall:

<u>Youth</u>

- Demographics: age, sex/gender, race, ethnicity, route of infection, education, living arrangements, orphan status, family resources
- Scores on screening mental health measures
- Scores on baseline questionnaires
    - Disclosure
    - Self-reported ART adherence
    - Barriers to adherence
    - Gender-based Violence (questionnaire about violence between male and female partners)
    - Gender Roles (questionnaire about relations between men and women)
    - AIDS-Related Stigma Scale
    - AIDS Risk Behavior Assessment
    - Children's Report of Parenting Behavior Inventory
- HIV-1 RNA (viral load)
- Current ARVs and psychiatric medications

Caregivers

- Demographics: age, sex, race, ethnicity, education, family resources, relationship to youth
- Scores on baseline questionnaires
    - AIDS-Related Stigma Scale
    - HIV attitudes and support
    - Barriers to adherence
    - HIV knowledge
    - CBCL caregiver report of youth behavior

5.2.3 Study status

Off-study reasons will be summarized by arm separately for youth and caregivers.

5.2.4 Study treatment

The following will be summarized by arm separately for youth and caregivers:

Youth

- Number of treatment sessions attended
- Number attending booster session
- Scores on mental health scales (PHQ-9, GAD-7, UCLA PTSD-RI) and other questionnaires at
    - IPL
    - 6-months
- Summary of "Risk for Contamination" responses
- Summary of TI-CBT fidelity
- If applicable, reasons for not receiving allocated intervention.

Caregivers

- Number of treatment sessions attended
- Number attending booster session
- Scores on questionnaires at
    - IPL
    - 6-months
- Summary of "Risk for Contamination" responses
- Summary of TI-CBT fidelity
- If applicable, reasons for not receiving allocated intervention.

5.2.5 Adverse events and deaths

Note: Reportable adverse events are described in protocol section 7.2 and include the following: Grade 2 or higher suicidal ideation or attempt, Grade 3 or higher psychiatric disorders, and Grade 3 or higher insomnia. All adverse events meeting criteria for expedited adverse event reporting, per Section 7.3.2 of the protocol.

- Summary of all new, post-entry reportable adverse events by arm
- Summary of reported social harms by arm
- Listing of deaths, if any, including treatment arm and cause
- Summary of the secondary safety analysis (see Section 4.2.2 for analytic details)

5.2.6 Efficacy

See Sections 4.2.1 and 4.2.2 for analytic details

- Summaries of mental health measures (individual measures and composite) by arm
- Summaries of composite mental health measure by sex/gender by arm
- Primary efficacy analyses – summary of analyses comparing treatment arms on mental health scores, including 95% CIs and p-values for differences in scores
- Secondary efficacy analyses – summary of analyses comparing treatment arms on mental health scores, adherence measures, and viral load
- Supplementary efficacy analyses – summary of univariable and final multivariable multi-level models, analysis of two-component mental health composite, and per-protocol analysis
- Supplementary missing data analyses – summary of loss to follow-up results if more than 10% of participants drop out before 6 months or if the arms differ significantly in rates of loss to follow-up
- Supplementary unscorable data analyses – summary of baseline characteristics between scorable and unscorable youth, and treatment effects by strategy (complete case or imputed)