Primary Statistical Analysis Plan

Version 1.0

For NICHD P1081

(Protocol Version 3.0 with LOA #1, #2, #3, #4, and #5)

A Phase IV Randomized Trial to Evaluate the Virologic Response and Pharmacokinetics of Two Different Potent Regimens in HIV Infected Women Initiating Triple Antiretroviral Regimens between 20 and 36 Weeks of Pregnancy for the Prevention of Mother-to-Child Transmission

ClinicalTrials.gov Identifier: NCT01618305

02/28/2019

Created by:

Center for Biostatistics in AIDS Research

Harvard T.H. Chan School of Public Health

Filename: *P:\documentation\analysisplans\NICHD_P1081_PRIMARY_SAP_v1.0_FINAL.docx*
Directory mapping: *Directory P maps to \\disney\home\impaact\P1081*

1 Introduction

1.1 Purpose

This Primary Statistical Analysis Plan (SAP) describes the primary and secondary outcome measures and additional outcome measures of the NICHD P1081 study that will be included in the primary manuscript, and which address, at a minimum, the primary and key secondary objectives of the study. 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. It also describes the analyses 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 Week 24 study visit, all queries have been resolved, and the study database closure/data lock has been completed.

Outlines of analyses for objectives and outcome measures not included in the Primary SAP will be provided in a separate SAP at a later date.

1.2 Key SAP Updates

The table below summarizes major revisions to the SAP that resulted in a version change. Updates will be made as necessary.

In the event of revisions after Version 1.0, bolded text will be used throughout the SAP to indicate major changes.

Version	Changes Made	Rationale	Effective Date
0.3	Original version pre-specifying details of the interim analyses for DSMB review	Pre-specify the analysis plan before the first interim efficacy analysis	03/29/2017
0.5	Converted to the new statistical analysis plan template; added more details, modifications based on DSMB and reviewer requests, and additional analyses to be performed for the final analysis;	Study has passed the primary completion date and is approaching the final study visit. Sent to writing team for review prior to sending preliminary working draft report for late breaker abstract preparation.	12/04/2018

1.0	Incorporated writing team comments and requests for additional analyses.	Final version used for Primary Final Analysis Report	02/28/2019
1.0	Removed “DRAFT” watermark	Editorial change	6/10/2019

2 Study Overview

2.1 Study Design

P1081 is a multicenter, two arm, randomized, open-label trial comparing the ability to achieve virologic suppression at delivery, tolerability, and safety in HIV-1 infected pregnant women with a gestational age between 20 and 36 weeks who are antiretroviral naïve or have received short-course zidovudine (for a maximum of 8 weeks) only for prevention of mother-to-child transmission (PMTCT) in previous pregnancies, and their infants.

P1081 has a target enrollment of 334 evaluable mother-infant pairs (approximately 167 per treatment arm), which is projected to require enrolling approximately 394 mother-infant pairs. Mothers (and infants) will be randomized in equal proportions to one of the two treatment arms. The randomization will be stratified by gestational age at entry (20- <28 weeks, 28-<31 weeks, 31-<34 weeks, and 34-<37 weeks) and by whether the mother will use lamivudine/zidovudine or an alternative, locally supplied nucleoside reverse transcriptase inhibitor (NRTI). The rationale for stratifying the randomization by gestational age is that women who enter the study later in gestation will be less likely to achieve the desired viral load decrease compared with women who enroll earlier in gestation. The rationale for stratifying the randomization by the chosen NRTI backbone is ensure balance in the two treatment arms, in case of unforeseen differential effects on viral load, tolerance, safety, or pregnancy outcomes.

Participants will be randomized to either arm A (lamivudine/zidovudine + efavirenz) or arm B (lamivudine/zidovudine + raltegravir) and receive their study drugs antepartum. During active labor, all participants will continue to receive study drugs. In addition, in place of the oral fixed dose combination of lamivudine/zidovudine, participants may receive intravenous zidovudine, other dosing regimens of oral zidovudine, or lamivudine and/or additional drugs during labor, according to local standard of care/guidelines.

Infants will receive ARV treatment according to specific local guidelines.

All women will receive their randomized study regimen from study entry through delivery. Women who meet local guidelines for receiving antiretroviral therapy (ART) will continue triple ART after delivery according to local guidelines.

2.2 Hypotheses

Primary Hypothesis:

1. Efavirenz and raltegravir are effective, safe and tolerable as part of HAART regimens to be used in late pregnancy when rapid viral load suppression is for PMTCT of HIV.

Secondary Hypotheses:

1. Efavirenz-based triple ARV regimens will decrease the level and infectivity of plasma and cell-associated virus more rapidly (by 1 week of ART) compared to II-based triple ARV regimens.
2. Transmitted HIV drug-resistance among women will be prevalent at 10-15% of the population. Transmitted resistance will be associated with delayed decay of plasma HIV-1 RNA levels compared to women without primary resistance, and, when ART is stopped with further selection of resistance (especially selection of lamivudine and/or NNRTI resistance).

2.3 Study Objectives

This Primary SAP addresses the primary and key secondary objectives listed in the study protocol. Other study objectives listed in the protocol will be addressed in subsequent analysis plans. All objectives outlined in the protocol are listed below. Those that are not covered by the Primary SAP are indicated as "**Not addressed in this Primary SAP**".

Note: One secondary objective, to "Compare decay of plasma HIV-1 infectivity between the treatment regimens", is no longer relevant. This objective was applicable under Version 2.0 of the protocol when it was a three-arm study with an arm containing a Protease Inhibitor (PI). Because this arm was dropped in Version 3.0 of the protocol, the objective is no longer valid. Therefore, the infectivity assays will not be run and this objective will not be addressed by this, or any subsequent analysis of P1081 data.

2.3.1 Primary Objectives

1. To compare the ability of two triple ARV regimens (one containing efavirenz and the other raltegravir) begun during the third trimester of pregnancy to achieve a viral load of <200 copies/mL at the time of delivery. [Protocol Objective 2.1.1]
2. To compare the safety and tolerability of two triple ARV regimens (one containing efavirenz and the other raltegravir) begun during the third trimester of pregnancy. [Protocol Objective 2.1.2]

2.3.2 Secondary Objectives

1. To compare the kinetics of viral decay between the treatment regimens [Protocol Objective 2.2.1]
    a. Compare decay of plasma and vaginal HIV-1 RNA and DNA between the treatment regimens.

   b. Compare decay of plasma HIV-1 infectivity between the treatment regimens. [**Note: because the PI arm was dropped from a prior version of this study, this objective is no longer relevant; the infectivity assays will not be run and this objective will not be addressed.**

2. To compare infant outcomes including stillbirth, premature birth, low birth weight, perinatal HIV transmission and to compare (in HIV-infected infants) drug resistance between the two treatment regimens.
3. To assess the baseline prevalence and selection of HIV-1 drug-resistance to the study drugs, using standard genotyping and ultrasensitive genotyping methods.

2.3.3 Exploratory Objectives

1. To describe the population pharmacokinetic (PK) parameters of efavirenz and raltegravir during the third trimester of pregnancy and postpartum using sparse sampling and to evaluate potential relationships between PK parameters, pharmacogenomics and viral load changes. [**Not addressed in this Primary SAP**]
2. To describe the maternal vaginal and infant nasopharyngeal and oropharyngeal microbiome environment and the potential association with adverse outcome in HIV exposed uninfected children. [**Not addressed in this Primary SAP**]

2.4 Overview of Sample Size Considerations

Justification for the proposed sample size is given in Protocol Section 8.4.1. In brief, the target sample size of 334 evaluable mother-infant pairs was chosen to provide 80% power to detect an important difference (an absolute difference of ≥15% was deemed clinically important to detect) between the two treatment arms in the primary efficacy outcome measure (defined in Protocol Section 8.2.1), with a two-sided Type I error rate (α) of 0.05, and allowing for interim efficacy analyses and non-evaluable women. Allowing for 5% of women to be non-evaluable for the primary efficacy outcome measure and another 10% of enrolled women to have genotypic resistance to study drug(s) at entry (see Protocol Section 1.6 for details regarding this assumption), an overall non-evaluability rate of 15% was assumed. Under this assumption, a target accrual of 394 enrolled women was proposed to achieve the target evaluable sample size of 334 women.

2.5 Overview of Formal Interim Monitoring

2.5.1 Ongoing Team Monitoring

The core protocol team will have regular conference calls to ensure that its members are aware of ongoing issues concerning the conduct of the study and will review reports about the status of the study on a monthly basis (the frequency may be decreased if the study team deems this appropriate). These will include reports on accrual, baseline characteristics, AEs, specimen completeness, and the proportions of women who are non-evaluable for the primary outcome measure or are found to have had genotypic resistance to any of the study drugs at screening (as defined in protocol Appendix IV). These reports will present results that are pooled across the randomized treatment arms and not broken out according to arm.

The core protocol team will monitor safety closely. A summary of maternal and infant AEs will initially be generated monthly to help identify possible safety issues early on. The frequency of these reports may be decreased to bimonthly or quarterly if no significant safety concerns are identified.

Accrual to this study will be monitored by the NICHD and protocol co-chairs in accordance with standard operating procedures. Also, the team will monitor site protocol activation of the African sites to ensure that the number of sites participating is sufficient to complete the accrual in a timely fashion. If accrual is not adequate to meet the enrollment goals specified in protocol Section 8.4.2, the team will identify the reasons for lack of accrual and possibly amend the protocol accordingly.

2.5.2 DSMB Reviews

This study will also be monitored by a NIAID-sponsored Data and Safety Monitoring Board (DSMB). The DSMB will review information concerning accrual, characteristics of participants, quality and completeness of data and specimen collection, retention, AEs, and the proportions of women who are non-evaluable for the primary outcome measure or who are found to have had genotypic resistance to any of the study drugs at screening (as defined in protocol Appendix IV) at least annually after the first woman is randomized.

Two interim efficacy analyses will be conducted when data on the primary outcome measure are available for approximately one third and two thirds of the planned enrollment. Under the accrual assumptions in the protocol, we anticipate that these interim analyses would be reviewed approximately one year and two years after the first enrollment to Version 3.0. The interim efficacy analysis schedule may be modified if accrual assumptions turn out to be inaccurate or if recommended by the DSMB.

The interim efficacy analyses will be based on comparison of the primary outcome measure between treatment arms, as described in Section 2.8 below. The Haybittle-Peto stopping boundary will be used as a guideline for considering a recommendation of early stopping. This guideline requires a p-value <0.001 at an interim analysis for early stopping to be considered.

To assist with decisions about recommending early stopping for lack of benefit (futility), conditional power and predicted interval analyses will be presented to the DSMB. The conditional power analysis will assess the power to detect the hypothesized treatment differences specified in protocol Section 8.4 upon continuation, conditional on the data observed so far. The predicted interval analysis will provide information on effect size estimates and potential improvement in precision upon continuation, under various assumptions regarding the data yet to be collected (e.g., that hypothesized treatment differences are true, that the observed trend continues, that the null hypothesis is true, and under best-case and worst-case scenarios). As a non-binding guideline for lack of benefit (futility), if the conditional power is low, say less than 20%, and the projected improvements in precision of effect estimates upon continuation are small, a recommendation of early termination may be considered. However, due to the lack of and need for efficacy and safety data for potent ARV regimens in the P1081 study population, the protocol team requests that the DSMB consider both the results of the above analyses and other factors that may argue for or against continuation (including whether there are safety or ethical concerns, the accrual rate, information to be gained from secondary objectives and sub-studies, new internal or external scientific information, and the existence/progress of other trials addressing the study questions), in deciding whether to recommend early stopping.

Although a recommendation for early termination would be based primarily on the primary efficacy analysis, consideration should be given to the consistency of effects seen on the primary and secondary efficacy outcome measures. Strong evidence of a difference in the primary outcome measure favoring one arm, but with evidence favoring the other arm with an important secondary efficacy outcome measure, might support the continuation of both arms. However, a significant difference between arms with respect to a secondary efficacy outcome measure, in the absence of strong evidence of a difference with respect to the primary outcome measure, would not be grounds for early stopping of an arm.

3 Outcome Measures

This Primary SAP includes analytic detail for all primary and secondary outcome measures that will be included in the Primary Final Analysis Report or submitted to ClinicalTrials.gov (regardless of the timeline for reporting).

Note: As noted above, one secondary objective, to “Compare decay of plasma HIV-1 infectivity between the treatment regimens”, is no longer relevant and the infectivity assays will not be run. Therefore, there are no outcome measures associated with this objective.

3.1 Primary Outcome Measures

3.1.1 Efficacy

Primary Efficacy Outcome Measure [Protocol Objective 2.1.1]: Plasma HIV-1 viral load <200 copies/mL at the delivery visit (or if there is no viral load measurement at the delivery visit, viral load <200 copies/mL within 3 weeks prior to delivery).

3.1.2 Safety and Tolerability

Primary Safety Outcome Measure [Protocol Objective 2.1.2]: Occurrence of at least one “new” adverse event of Grade ≥3 as defined in the DAIDS Toxicity Grading Table through Week 24 postpartum. This analysis will be done separately for each of women and infants. “New” events for women are signs and symptoms, hematologies, chemistries, and diagnoses that occur on or after randomization (or increase in Grade after randomization). For infants, “new” events are those adverse events that occurred on or after birth.

Primary Tolerability Outcome Measure [Protocol Objective 2.1.2]: Permanent discontinuation of efavirenz or raltegravir (whichever was assigned) prior to labor and delivery for any reason (including loss to follow-up) will be considered a treatment failure in this analysis (note: switching any of the NRTIs with continuation of efavirenz or raltegravir will not be considered a treatment failure). Temporary holds of efavirenz or raltegravir will not be considered a treatment failure.

3.2 Secondary Outcome Measures

3.2.1 Secondary Efficacy Outcome Measures

The following are secondary efficacy outcome measures:

<u>Viral Load at Delivery</u> [Protocol Objective 2.1.1]

- Virologic suppression to below the lower limit of quantification (LLQ) of the assay at, or within 21 days prior to, delivery.
- Almost all women had their plasma HIV-1 RNA viral load at delivery (or last viral load up to 21 days prior to delivery) measured using an assay with an LLQ=40. For these women, a viral load ≤40 with a censor code indicating that the viral load was below the LLQ will be considered successes. Others will be considered failures.
- For women who had their viral load measured using an assay with an LLD≠40, the outcome measure will be defined as follows:
  - For women who had their viral load measured using an assay with a LLQ<40, all women will be considered successes who had a viral load <40, regardless of whether the censor code indicates that the measured viral load was below the LLQ. If the measured viral load is ≥40, they will be considered failures.
  - For women who had their viral load measured using an assay with a LLQ>40, the outcome measure will be defined in two ways.
    - In the primary analysis of this outcome measure, they will be considered successes if their viral load was ≤LLQ and the censor code indicates that the value was below the LLQ, regardless of what the LLQ for that assay was. Otherwise, they will be considered failures.
    - In a sensitivity analysis of this outcome measure, all women who had a LLQ >40 will be considered failures, regardless of what the measured viral load and censor code were.

<u>Composite Efficacy Outcome Measure</u> [Protocol Objectives 2.1.1 and 2.1.2]

A key secondary outcome measure will be a composite outcome measure that combines efficacy and tolerability (Protocol Objectives 2.1.1 and 2.1.2). Specifically, this composite outcome will be a binary outcome measure of (1) a successful viral load (plasma HIV-1 RNA) decrease from entry to study week 2 (day 11-17) and viral load <1,000 copies/ml at all time points after 4 weeks on study drugs, until delivery; and (2) tolerability (remaining on the assigned study regimen). The viral load decrease and tolerability components of the composite outcome measure will be defined as follows:

- Rapid viral load decrease for women who deliver after 4 weeks on study drugs: A successful viral load decrease is defined as having both (i) a plasma HIV-1 RNA level ≥2.0 log10 below baseline or <200 copies/mL at Week 2 (day 11-17 after initiation of treatment) and (ii) a plasma HIV-1 RNA level <1,000 copies/mL at all time points after 4 weeks on study drugs, until delivery.

- Rapid viral load decrease for women who deliver before or at 4 weeks on study drugs: A successful viral load decrease is defined as a plasma HIV-1 RNA level ≥2.0 log10 below baseline or <200 copies/mL at Week 2 (day 11-17 after initiation of treatment).
- Tolerability: Permanent discontinuation of efavirenz or raltegravir prior to delivery for any reason (including loss to follow-up) will be considered a treatment failure in this analysis (note: switching any of the NRTIs with continuation of efavirenz or raltegravir will not be considered a treatment failure). Temporary holds of efavirenz or raltegravir will not be considered a treatment failure.

The baseline value will be the value obtained at the study entry visit. If this value is not available, then the baseline value will be the screening value. If a woman has more than one viral load measurement within a visit window, the earliest measurement will be used in the analysis.

3.2.2 Kinetics of Viral Decay Outcome Measures

Kinetics of Viral Decay [Protocol Objective 2.2.1]

- HIV-1 RNA and DNA viral load in maternal blood and vaginal swabs (only blood plasma RNA viral load will be analyzed in the primary analysis report; vaginal swabs and blood plasma DNA viral load will be assayed at a later date) at Week 4 and 6 from initiation of treatment.
    - For viral load, the outcome measures are viral load <200 copies/mL and <LLQ. This outcome will be analyzed twice; once at Week 4 and once at Week 6. The viral loads closest to 28 days (but within 24-32 days) and 42 days (but within 38-46 days) from initiation of treatment will be used for Week 4 and Week 6 respectively. The viral load obtained closest to the target date will be used for each visit. In the event that there are multiple viral loads that were obtained the same amount of time from the target visit date, the earliest of the viral loads will be used.
    -

- Change on a $\log_{10}$ scale in HIV-1 RNA and DNA viral load from entry (or screening if there is no entry viral load) to each time point prior to delivery.
    - The viral load at each time point will be the viral load closest to the time point and within the allowable visit window given in the Schedule of Evaluations (Protocol Appendix I). If there are multiple viral loads that were taken an equal amount of time from the target date for that visit, the earliest will be used.
    - Log_{10} change from baseline to Week X will be calculated as Log_{10}(Week X viral load) – Log_{10}(Baseline viral load).

- Infectivity of plasma during the initial 2 weeks of Antiretroviral Therapy (ART) is **not addressed**; as previously described, this outcome measure is no longer relevant and the assays will not be done).

3.2.3 Infant Outcomes:

Adverse Pregnancy Events [Protocol Objective 2.2.2]:

The following adverse pregnancy outcomes will be examined individually and analyzed both individually and in combination:

- Stillbirth/fetal demise
- Premature birth (<37 weeks gestation at delivery)
- Low birth weight (<2500 grams)

For each analysis, the unit of analysis will be the pregnancy (i.e. mother-infant pair). All pairs where at least one delivery occurred on-study will be eligible for each analysis. If a mother carrying twins had discordant pregnancy outcomes, the mother will be considered an event if either of the twins had that outcome. A sensitivity analysis will include only M-I pairs where the mother was evaluable for the primary efficacy analysis.

In addition to the analyses for each individual adverse pregnancy outcome, a composite analysis will be performed. A M-I pair will be considered an event if an infant (or at least one of the infants in the case of twins) that was delivered on-study experienced at least one of the listed adverse pregnancy events. The analysis population will be the same as for the individual analyses. Similarly, a sensitivity analysis that includes only M-I pairs where the mother was evaluable for the primary efficacy analysis will be performed.

NOTE: If numbers permit, both the individual analyses and the composite analysis will be repeated, with more extreme definitions for premature birth (<34 weeks gestation at delivery) and low birth weight (<1500 grams).

Infant HIV-Infection [Protocol Objective 2.2.2]

- Infant HIV-infection status will be analyzed. Infants who are classified as “Infected” or “Probably infected, based on best available data” will be considered infected. Infants who are classified as “Uninfected” or “Negative, based on best available data” will be considered uninfected.

HIV-1 Drug Resistance in HIV Infected Infants [Protocol Objective 2.2.2]

- HIV-1 drug resistance mutations (as defined in Protocol Appendix IV) will be examined at the time of detection of infant HIV infection. These samples will not be assayed in time to include in the primary analysis report, so they will be included in a subsequent analysis report.

3.2.4 HIV-1 Drug Resistance Outcome Measures

HIV-1 Drug Resistance [Protocol Objective 2.2.3]

- HIV-1 drug resistance mutations (as defined in Protocol Appendix IV) will be examined at the following timepoints:
    - Screening (among women)
    - Between 2-4 weeks postpartum (among women who have stopped ART)
    - At the time of inadequate virologic response as defined in Protocol Section 6.2.9
- All women who have a resistance result for at least one class (RT and/or IN) of ARVs at that time point (screening, postpartum, and/or at the time of inadequate virologic response as defined in the Protocol) will be evaluable at that time point. Because IN resistance is expected to be rare, RT and IN resistance will be analyzed and described as separate outcomes.
- Only screening samples among women will be assayed in time to be included in the primary analysis report. All postpartum samples for women will be assayed and analyzed in a subsequent analysis report.

3.2.5 Post Hoc Outcome Measures

Composite Ranked Endpoint of Pregnancy Outcomes and Safety Events [Protocol Objectives 2.1.2 and 2.2.2]

- The following outcomes will be ranked from worst (top) to best (bottom):
    - Infant or Maternal death
    - Fetal death (≥ 20 weeks gestation)
    - Infant HIV infection
    - Extremely and very early preterm (<34 weeks gestational age at delivery)
    - Congenital anomaly
    - Extremely low birthweight (<1500 grams)
    - Preterm delivery (34 to <37 weeks gestational age at delivery)
    - Hospitalization
    - Low birthweight (1500-<2500 grams)
    - Grade 3 or 4 adverse event (maternal or infant)
    - None of the above
- The composite ranked outcome will then be analyzed.

Time-to Viral Suppression

- Time-to viral load<200 copies/mL will be calculated as the number of days from entry. Because viral suppression is measured at distinct visits, true suppression occurs sometime between the last visit where viral load≥200 copies/mL and the first visit where viral load<200 copies/mL. Therefore, we calculate an interval censored time-to-event as follows:
    - The lower bound of the interval will be the number of days from entry to the last observed patient visit on or before delivery where the participant did not have a viral load<200 copies/mL.
    - The upper bound of the interval will be the number of days from entry to the first visit on or before delivery at which the participant had a viral load<200 copies/mL.
    - Participants who deliver or go off-study prior to delivery and prior to achieving viral load<200 copies/mL will only have a lower bound (i.e. these participants will be censored at their last contact on or before delivery).

3.3 Exploratory Outcome Measures

All exploratory outcome measures will be analyzed at a later date. The samples needed to address these outcomes will be assayed in batch at a later date; thus, the analysis of these objectives and the outcome measures used will be described in detail in a subsequent analysis plan, and are not addressed in this Primary SAP. These outcome measures will not be submitted to CT.gov.

4 Statistical Principles

4.1 General Considerations

- All participants randomized will be considered eligible for inclusion in analyses, regardless of whether they are later determined to be ineligible for the protocol. Specific analysis populations may have additional requirements for eligibility.
- Additions of or changes in outcome measures that are identified after analysis has begun will be identified as post hoc.
- Unless otherwise noted, baseline/entry values refer to the value closest to and on or before randomization.
- Visit windows used in analysis will be the expected date of the visit +/- 3 days (e.g. a Week 4 analysis visit would consider values observed from day 25-day 31). If there are multiple values within an analysis window, the value in the database that was observed closest to the expected date of the visit will be used. In the event that multiple values are equidistant from the expected visit date, the earliest of these observed values will be used.
- Because this is a randomized clinical trial, no statistical comparisons across groups for baseline characteristics are planned.
- For interim efficacy reviews, the Haybittle-Peto stopping guideline will be used. P-values<0.001 will be considered statistically significant, and may inform the DSMB's decision to stop the trial. For safety analyses at interim reviews, the nominal P-value will be considered significant if P<0.05. Because the Haybittle-Peto guideline is used for each interim efficacy analysis, no adjustment for alpha-spending will be used. In final analysis, all efficacy, safety, and tolerability comparisons will be considered statistically significant if P<0.05.
- For Phase III and pivotal Phase II and IV studies, NIH requires primary analyses of treatment comparisons to be summarized by sex and by race and treatment interactions with sex and race to be tested. For NICHD P1081, the primary analyses of treatment comparisons will be summarized by race and treatment interactions with race will be tested (analyses by sex are not possible because NICHD P1081 does not enroll men). These analyses are required so do not represent multiple comparisons and are presented in the primary study analysis regardless of power issues.

4.2 Analysis Populations

4.2.1 Eligibility Violations and Exclusions

Two women who were randomized to NICHD P1081 were later discovered to have received ARVs prior to study entry (participants are required to be ARV naïve, with at most 8-weeks short-course ZDV for the prevention of transmission in a previous pregnancy). These women were kept on study per the Intent-to-Treat (ITT) principle, and are eligible for inclusion in all analyses (provided they meet all other eligibility requirements for that analysis).

One infant who was delivered on-study (as a live birth) was noted to be part of a multiple gestation. However, the M-I pair was randomized as a singleton birth. Upon querying the site, the woman was determined to have had a spontaneous abortion that resulted in fetal demise of the second infant at approximately 18 weeks gestation. Because eligible women for the study must have a gestational age of at least 20 weeks at entry, this spontaneous abortion outcome occurred prior to study entry. Because the twin's outcome occurred prior to study entry, this woman and her live-birth infant will be considered to be a singleton M-I pair in all analyses of outcome measures where a comparison is performed (i.e. a P-value is produced); the twin and its outcomes will be excluded from all such analyses. However, due to data discrepancies that would result from reporting a singleton birth from a multiple gestation, the twin that was a fetal demise at 18 weeks was retroactively enrolled so that the outcome may be reported in descriptive tables.

4.2.2 Primary Efficacy Population

The primary efficacy outcome is viral load at (or within 21 days prior to) delivery. Therefore, eligible women will be those women who remain on-study through delivery and have a delivery CRF in the database. Evaluable women will be those with either a screening or entry viral load ≥200 copies/mL, a valid HIV-1 RNA viral load at (or within 21 days prior to) delivery, and a valid genotypic resistance result for all study ARV classes that indicates no known resistance mutations (a detailed list of protocol-specified resistance mutations is given in Protocol Appendix IV). Women who have known resistance or a missing resistance result for at least one study ARV class will be excluded from the primary efficacy analysis, but will be included in sensitivity analyses (further details in Section 4.3 Analysis Approaches).

4.2.3 Primary Safety Population

All women and infants enrolled will be considered eligible for inclusion in their respective primary safety analyses (the primary safety analysis will be done separately, both among women and among infants). Women who receive at least one dose of their assigned study drug, and their infants, will be evaluable for their respective primary safety analyses.

4.2.4 Primary Tolerability Population

All women enrolled will be considered eligible for inclusion in the primary tolerability analysis. Women who received at least one dose of their assigned study drug will be considered evaluable. Because only women receive study drug, infants will not be included in this population.

4.3 Analytic Approaches

4.3.1 Primary Analyses

For each of the primary efficacy, safety, and tolerability analyses, the comparison between arms of the proportion of participants who meet each respective outcome will be assessed using a Cochran-Mantel-Haenszel (CMH) test stratified by the following gestational age strata: 20-<28 weeks, 28-<31 weeks, 31-<34 weeks, and 34-<37 weeks. Further considerations unique to each analysis are described in following sections.

Primary Efficacy Analysis

The primary efficacy analysis will be a comparison of the proportion of women who have viral load <200 ml at (or within 21 days prior to) delivery. This analysis will be performed three ways:

A. The primary analysis will exclude women who have genotypic resistance to any study drugs at screening or do not have complete resistance results.
B. A secondary analysis will include these women, to compare the two real-world strategies of starting therapy with either Arm A or Arm B and possibly switching ARVs when the resistance test results become available, subject to the potential biases described in protocol Section 8.1.
C. For the final analysis only, a sensitivity, “all-comers” analysis will include all include all women who have a viral load measurement at or within 21 days prior to delivery, regardless of genotypic resistance results or baseline viral load. This analysis is subject to the same biases as the secondary analysis above.
D. For interim analyses only, if there are substantial numbers of missing raltegravir resistance results (e.g., >10% missing) because integrase resistance testing is not standard procedure at some labs, a sensitivity analysis will repeat the primary analysis after including the women who are missing raltegravir resistance results but are otherwise evaluable, and these women will be counted as having no raltegravir resistance (based on two studies showing a very low prevalence of genotypic resistance to integrase inhibitors in Brazil).

For interim analyses only, the above efficacy analyses will be restricted to evaluable women who had an estimated delivery date 4 weeks or more before the data freeze date, to avoid overrepresentation of preterm deliveries that occur right before the data freeze date. A sensitivity analysis will also be conducted in which the above analyses are repeated after including all evaluable women who delivered up to the data freeze date. For the final analysis (when all women will have already delivered), no restriction on estimated delivery date will be used in the efficacy analyses.

Sensitivity analyses will be conducted as needed to assess the potential impact of missing data on the conclusions of the study. Of primary concern are missing viral load measurements at delivery for women who have achieved a successful viral load decrease at time points at which measurements are available. In sample size calculations, it was assumed that approximately 5% of women would be missing their viral load measurement at delivery (or within 21 days prior). Therefore, the proportion of women who have a viral load measurement at delivery (or within 21 days prior to delivery) will be examined, and sensitivity analyses will be done if <95% of eligible women have a viral load measurement (i.e. if >5% of women do not have a delivery viral load measurement). The sensitivity analyses will be done in two ways: (a) as an extreme, by assuming that a missing viral load measurement at delivery would have shown successful or unsuccessful viral load decrease in a way that would minimize the difference between randomized groups, and (b) more plausibly, by assuming that a missing viral load measurement at delivery would have shown an unsuccessful viral load decrease with probability equal to the estimated probability of an unsuccessful viral load decrease at delivery among women in the same group who had that evaluation and had a successful viral load decrease at other evaluations prior to delivery.

If the proportion of women who are missing their viral load measurement at delivery is much larger than expected, multiple imputation analyses may also be warranted. Therefore, if >10% of women are missing their viral load measurement at delivery (i.e. <90% of eligible women have a viral load measurement at delivery) multiple imputation will be performed. The primary efficacy analysis will be replicated including women who were missing their viral load at delivery (but excluding those who were non-evaluable for other reasons). These women who have a missing delivery viral load will have that viral load imputed via logistic regression. Thirty imputations will be performed and each imputed dataset will be analyzed individually as described in the primary analysis above. The CMH statistic that is calculated in each of these thirty analyses will then be transformed and pooled to generate an overall CMH statistic and associated P-value (Note: Because the chi-square distribution is highly skewed for smaller degrees of freedom, obtaining a combined result of the CMH test from multiple imputations requires a transformation to normalize the CMH statistic. In this analysis, the Wilson-Hilferty transformation will be used for this purpose.

Since both the randomization and primary analysis are stratified by gestational age at enrollment (because women who enrolled later in gestation would be less likely to achieve virologic suppression at delivery), a test for a treatment-by-gestational age stratum interaction will be implemented. This test will be performed via logistic regression among the primary efficacy analysis population, and will include treatment (efavirenz vs raltegravir), gestational age stratum (20-<28 weeks vs. 28-<31 weeks vs 31-<34 weeks vs. 34-<37 weeks), and the treatment-by-gestational-age-stratum interaction as predictors. The primary outcome measure (HIV-1 RNA viral load <200 copies/mL) will be the response variable. If the model does not converge, for example due to a small or zero event-rate in one arm/stratum combination, then adjacent gestational age strata will be combined. They will be combined first by trimester (20-<28 weeks vs. 28-<37 weeks). If this model also fails to converge, additional combinations will be made by combining adjacent strata. These combinations will attempt to keep the number evaluable in each new stratum balanced, and may combine adjacent strata when one stratum (or more) have a low or zero event rate.

If there is evidence of a significant treatment-by-gestational age stratum interaction, stratum-specific estimates of the primary efficacy outcome measure for each treatment arm will be provided for the four gestational age strata used for the randomization, for each of the three efficacy analysis populations: the primary efficacy analysis population (Analysis A above), the population that includes those with genotypic resistance (Analysis B above), and the population that additionally includes those with baseline/screening RNA<200 copies/mL (Analysis C above).

Primary Safety Analysis

The primary safety analysis will be a comparison of the proportion of participants who experienced at least one grade 3+ adverse event on or after randomization. This comparison will be done twice; once for women and once for infants.

For women, this analysis will be replicated in additional sensitivity analyses. The safety sensitivity analysis will replicate the primary safety analysis, but will be censored at delivery; only events that occur on or before delivery will be considered (the primary safety analysis includes all events through the end of follow-up at 24 weeks postpartum). Both the primary and sensitivity safety analyses will also be replicated including only women who were evaluable for the primary efficacy outcome measure.

Primary Tolerability Analysis

The primary tolerability analysis will be a comparison of the proportion of women who discontinued their assigned treatment, for any reason, prior to delivery. This comparison is only performed among women because infants do not receive study treatment.

4.3.2 Secondary Analyses

Secondary Efficacy Analyses

- Viral load <LLQ at delivery
  - The proportion of women who achieve a viral load below the lower limit of quantification at (or within 21 days prior to) delivery will be compared between arms using a CMH test. Women who were evaluable for the primary efficacy analysis will be evaluable for this outcome measure. If there is a significant proportion of women who are non-evaluable due to missing genotypic resistance results, this analysis will be replicated in sensitivity analyses using the same approach as the primary efficacy analysis.
  - The analysis population, sensitivity analyses, and stratum-specific analyses of this outcome measure will be the same as for the primary efficacy outcome.
  - The above analyses will be replicated using a standardized LLQ. The definition of this outcome has been defined previously in Section 3 of this Primary SAP.

- Composite Efficacy Analysis
  - The composite efficacy analysis will be a comparison of the composite efficacy outcome measure between arms using a CMH test stratified by gestational age at entry. The gestational age strata will be the same as for the primary analyses.
  - Similarly, the composite efficacy analysis will be performed in the same three ways as the primary efficacy analysis. The interim analyses will also be restricted to women who had an estimated delivery date four weeks or more prior to the data freeze date. For final analysis, no restriction on estimated delivery date will be used.
  - Differential rates of non-evaluability or preterm delivery between study arms could lead to biased results; for example, an excess of preterm deliveries in one arm could lead to more women in that arm delivering before 4 weeks on study drugs and therefore not needing to maintain viral load <1,000 copies/ml after 4 weeks on study drugs, which could inflate the response rate in that arm. To address this concern, the rates of non-evaluability and preterm delivery will be compared between study arms,

additional analyses will compare the study arms with respect to each individual component of the composite outcome measure, and sensitivity analyses will be conducted to assess the potential impact of missing evaluations on the conclusions of the study (see Protocol Section 4.3.1 for details).

Kinetics of Viral Decay Analyses

- Maternal HIV-1 RNA blood plasma viral load at Weeks 4 and 6
  - The proportion of women who have viral load <200 copies/mL at Week 4 from initiation of treatment will be compared between the two arms via a CMH test stratified by gestational age at entry. This analysis will also be performed at Week 6.
  - The population for both analyses will include women who have a baseline viral load measurement; additionally, the comparisons at Week 4 and Week 6 will be performed among women who have a viral load result for that week.

- Log_{10} change in maternal HIV-1 RNA blood plasma viral load from entry to each time point prior to delivery
  - We will analyze this outcome measure descriptively at each time point. The median (Q1-Q3) change in Log_{10} viral load will be presented.
  - The population for this analysis will include women who have a baseline viral load measurement; additionally, the descriptive statistics at each week will be calculated among women who have a viral load result for that week.

- Log_{10} change in maternal HIV-1 RNA vaginal viral load, as well as both blood plasma and vaginal DNA viral load, will be analyzed as above among women. This analysis will be performed at a later date, when the virology results are available, and will not be included in the primary analysis report.

Infant Outcomes

- The proportion of infants who experience each of the following outcomes will be compared between the treatment arms. Because there are expected to be small numbers of each event observed, each comparison will be performed utilizing a Fisher's exact test.
  - The following events will be analyzed
    - Sillbirth/fetal demise
    - Premature birth will be assessed twice; once considering <34 weeks gestational age as a premature birth, and once considering <37 weeks gestational age as a premature birth
    - Low birth weight will be assessed twice; once considering a weight of <1500 grams as a low birth weight and once considering <2500 grams as a low birth weight.
    - Infant HIV infection status
  - The unit of analysis will be the M-I pair. If a woman had a multiple gestation, that M-I pair will be considered an event if at least one of the infants delivered experienced

that event. All women who delivered on study, and their infants, will be eligible for each analysis. Evaluable M-I pairs will be those that had a value recorded for that outcome measure. In the case of a multiple gestation, an M-I pair will be evaluable for analysis of each outcome measure if at least one of the infants had a recorded value for that outcome measure.

- Additionally, the proportion of M-I pairs that experience at least one adverse pregnancy outcome will be analyzed.
    - The analysis will be performed with two groups of events.
        - The first considers stillbirth, very premature birth (<34 weeks gestation), and very low birth weight (<1500 grams) as events.
        - The second considers stillbirth, premature birth (<37 weeks gestation) and low birth weight (<2500 grams) as events.
    - Both analyses will be performed using a Fisher's exact test. If the number of events permits, this test will also be replicated using a CMH test stratified by gestational age at entry (using the same gestational age strata as in the primary analyses).
    - Similarly, the unit of analysis will be the M-I pair. All women who delivered on study, and their infants, will be eligible for this analysis. Evaluable M-I pairs will be those where at least one infant in that gestation had a value recorded for at least one of the outcomes (birthweight, gestational age at birth, delivery outcome). A M-I pair will be considered an event if at least one of the infants born on study as part of that gestation had at least one of the outcomes.
- Among the HIV-infected infants, any HIV-1 drug resistance mutations defined in the Protocol (Appendix IV) that are identified at the time of first positive HIV-1 RNA test will be listed (the numbers are anticipated to be too small for any formal comparisons or summary statistics).
    - Note: Infant resistance data will not be available at the time of the primary final analysis, and thus will not be included in the primary analysis report.

HIV-1 Drug Resistance

- The proportion of women who show any resistance mutation defined in the Protocol (Appendix IV) on a sample taken at or before entry will be assessed descriptively, and a 95% confidence interval will be estimated using a Wilson score. The proportion of women who show resistance to each individual class of antiretrovirals (reverse transcriptase and integrase inhibitors) will also be assessed, and a 95% Wilson score confidence interval will be estimated. There will be no formal comparison between arms.
- A similar analytic approach will be used to estimate the proportion of women who have, or have developed, genotypic resistance to study drugs at Week 2-4 postpartum and at the time of inadequate virologic response.
- Note: The post-entry resistance data will not be available at the time of the primary final analysis, and thus will not be included in the primary analysis report.

4.3.3 Post Hoc Analyses

DOOR-type Composite Pregnancy Outcome

- The DOOR-type outcome analysis will be an analysis of the composite pregnancy outcome measure, the Composite Ranked Endpoint of Pregnancy Outcomes and Safety Events. This analysis will compare the two treatment arms via an ordinal logistic regression, adjusted for gestational age at entry (the same gestational age strata used as in the primary efficacy analysis will be included as a covariate).
- Sensitivity analysis: the above analysis will be replicated including only M-I pairs where the mother was evaluable for the primary efficacy outcome measure.

Time-to Viral Suppression

- Time-to viral suppression (HIV-1 RNA viral load<200 copies/mL) curves will be estimated using the Expectation Maximization Iterative Convex Minorant (EMICM) algorithm in SAS, using PROC ICLIFETEST.
  - The generalized log-rank statistic will be used to compare the treatment arms. The default weighting method of Sun (i.e. all weights are equal to one) will be used.
  - The variance and standard error of both the survival curves and the generalized log-rank statistic will be estimated using multiple imputation. The seed for the imputations will be the SAS date (20180) that corresponds to the date the latest version of the protocol was finalized (April 2nd, 2015), and 1000 imputations will be performed.
  - Point estimates and 95% confidence intervals for Q1, median, and Q3 time-to viral load<200 copies/mL will also be estimated for each treatment arm using the EMICM algorithm.
- Primary analysis: all women who received at least one dose of study drug and had baseline and/or screening viral load ≥200 copies/mL will be included.
- Sensitivity analysis: the above analysis will be replicated including only women who were evaluable for the primary efficacy analysis (Analysis A).

5 Report Contents

The Primary Final Analysis Report will contain the following sections:

- Introduction
- Background
- Consort Diagram
- Data Included in the Report and Definitions
- Statistical Considerations
- Accrual
- Eligibility Violations
- Baseline Characteristics
- Study Status of Women
- Safety among Women
- Tolerability among Women
- Pregnancy Outcomes
- Study Status among Infants
- Safety among Infants
- Infant HIV Infections
- Evaluability for the Primary Efficacy Outcome Measure
- Efficacy
- Post Hoc Analyses
- Summary