# A Phase 3 Study to Evaluate Weight Gain of ALKS 3831 Compared to Olanzapine in Adults with Schizophrenia

| | |
|---|---|
| **Unique Protocol ID:** | ALK3831-A303 |
| **NCT Number:** | NCT02694328 |
| **Date of Statistical Analysis Plan:** | 18 Sep 2018 |

# STATISTICAL ANALYSIS PLAN

**ALK3831-A303**

**NCT02694328**

| | |
|---|---|
| Study Title: | A Phase 3 Study to Evaluate Weight Gain of ALKS 3831 Compared to Olanzapine in Adults with Schizophrenia |
| Document Status: | Final Version 3.0 |
| Document Date: | 18 Sep 2018 |
| Based on: | ALK3831-A303 Protocol Amendment 5.0 Date: 18 Sep 2018 |
| Sponsor: | Alkermes, Inc.<br>852 Winter Street<br>Waltham, MA 02451<br>USA |

# TABLE OF CONTENTS

# LIST OF TABLES



# LIST OF FIGURES

# ABBREVIATIONS

| Abbreviation or Term | Explanation or Definition |
|---|---|
| AE | adverse event |
| AIMS | Abnormal Involuntary Movement Scale |
| ANCOVA | analysis of covariance |
| ATC | Anatomical Therapeutic Chemical [classification system] |
| BARS | Barnes Akathisia Rating Scale |
| BMI | body mass index |
| BPM | Beats per minute |
| CGI-I | Clinical Global Impression - Improvement |
| CGI-S | Clinical Global Impression - Severity |
| CI | confidence interval |
| CP | conditional power |
| C-SSRS | Columbia Suicide Severity Rating Scale |
| ECG | electrocardiogram |
| eCRF | electronic case report form |
| EOT | end of treatment |
| EPS | extra pyramidal symptoms |
| ESAP | Exploratory Statistical Analysis Plan |
| ET | early termination |
| FAS | full analysis set |
| GLMM | generalized linear mixed effect model |
| HbA1C | Hemoglobin A1c |
| IA | Interim Analysis |
| IAP | Interim Analysis Plan |
| IWQOL | Impact of Weight on Quality of Life |
| MCMC | Markov Chain Monte Carlo |
| MedDRA | Medical Dictionary for Regulatory Activities |
| MI | multiple imputation |
| MMRM | mixed model with repeated measurements |
| OLZ | olanzapine |

| Abbreviation or Term | Explanation or Definition |
|---|---|
| PANSS | Positive and Negative Syndrome Scale |
| PCS | potentially clinically significant |
| PK | pharmacokinetic |
| QTcB | QT corrected with Bazett formula |
| QTcF | QT corrected with Fridericia formula |
| SAE | serious adverse event |
| SAP | statistical analysis plan |
| SAS | Simpson-Angus Scale |
| TEAE | treatment-emergent adverse event |
| ULN | upper limit of normal |
| WHO | World Health Organization |

# 1. INTRODUCTION

This statistical analysis plan (SAP) describes the statistical methods to be used for analyses and data presentation for reporting efficacy and safety results for study ALK3831-A303. This document has been prepared based on Alkermes ALK3831-A303 Study Protocol.

## 1.1. Study Objectives

Primary objective:

- To evaluate weight gain of ALKS 3831 (a fixed-dose combination of olanzapine and samidorphan) compared to olanzapine in adults with schizophrenia

Secondary objective:

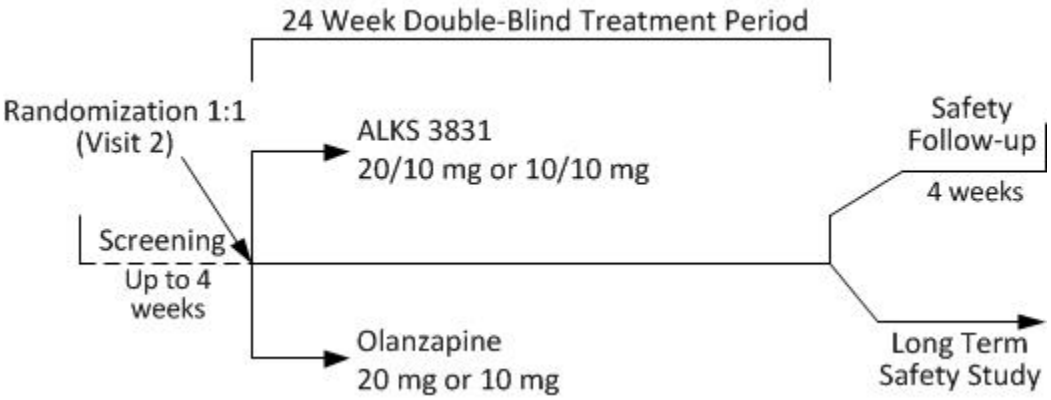
- To evaluate the safety and tolerability of ALKS 3831 in adults with schizophrenia

## 1.2 Summary of the Study Design

This is a Phase 3, multicenter, randomized, double-blind, study in adults with schizophrenia.

Subjects will be screened at Visit 1, up to 30 days prior to randomization. At Visit 2, eligible subjects will be randomized 1:1 to ALKS 3831 (olanzapine/10 mg samidorphan) or olanzapine treatment groups and receive study drug for up to 24 weeks. Subjects are not excluded from the study if they are taking other antipsychotic medication. However, they must be tapered off this medication and switched onto ALKS 3831 or olanzapine by the end of Week 2, Visit 4. This is an outpatient study in which subjects will visit the study site weekly for the first 6 weeks, then biweekly for the remaining 18 weeks. Study drug will be provided as coated bilayer tablets dispensed to the subjects in blister packs at each visit to be taken at home, one tablet by mouth each day, preferably at bedtime. For the first week, subjects will receive 10 mg olanzapine per dose: either ALKS 3831 10/10 (10 mg olanzapine/10 mg samidorphan) or OLZ 10 (10 mg olanzapine), according to treatment assignment. At the end of Week 1, the dose will escalate to 20 mg: ALKS 3831 20/10 [20 mg olanzapine/10 mg samidorphan] or OLZ 20). Following this increase, the dose may be decreased back to 10 mg at the end of Week 2, 3, or 4 according to investigator discretion. No further dose adjustments will be allowed beyond Week 4 and the dose will remain fixed for the remaining 20 weeks of the study. Body weight and metabolic parameters (including fasting triglycerides, cholesterol, and glucose) will be measured throughout the 24-week treatment period. Psychiatric symptoms will also be evaluated using Positive and Negative Syndrome Scale (PANSS) and Clinical Global Impression (CGI).
A schematic summarizing study design is presented below.

**Figure 1: ALK3831-A303 Study Design Schematic**




Subjects completing this study will be eligible to enroll in a long-term safety study (ALK3831-A304). Subjects not continuing in the long-term safety study or subjects prematurely discontinuing from the study will enter the 4-week safety follow-up period.

Premature discontinuation procedures are provided in Protocol Section 7.3. Randomized subjects who discontinue study drug but are willing to come in for further assessments will be asked to complete the early termination (ET) visit followed by 2 safety follow-up visits 2 and 4 weeks after the ET visit. They will then be asked to return to the study center for monthly visits until end of planned treatment, to collect the following information: weight, adverse events, and new antipsychotics.

# 2. SAMPLE SIZE CONSIDERATION

The initial target sample size is 200 randomized subjects per treatment group (400 in total). This sample size will provide at least 90% power to detect a significant difference in mean percent change in body weight at Week 24 of 4% or more (3.0% on ALKS 3831 versus 7% on olanzapine group with standard deviation of 9%) and a 13% difference in proportion of subjects with ≥10% weight gain at Week 24 (14% on ALKS 3831 versus 27% on olanzapine group).

An unblinded interim analysis is planned to be conducted to determine if the final sample size will be increased to 540 (270 subjects per treatment group), details in ALK3831-A303 Interim Analysis Plan.

# 3. DATA ANALYSIS

## 3.1. General Statistical Methodology

Baseline for the primary efficacy or safety analysis is defined as the last non-missing efficacy or safety assessment before the first dose of double-blind study drug.

In general, descriptive statistics: n, mean, standard deviation (SD), median, minimum, and maximum, for continuous variables and number and percentage of subjects in each category for categorical variables will be provided by treatment group for all variables.

The fasting lipids and fasting glucose laboratory parameters will be analyzed on a linear and log-transformed scale.

All statistical tests and confidence intervals (CIs), unless stated otherwise, will be 2-sided and will be set at an alpha level of 0.05.

To adjust for the unblinded interim analysis for sample size re-estimation, the CHW method will be applied to the primary analysis of the two primary endpoints only, with details in Section 3.8.2.1.1 and Section 3.8.2.2.1. It will not be applied to the sensitivity or additional analysis of the two primary endpoints, analysis of the key secondary, other, exploratory endpoints, or subgroup analyses.

All source data will be presented as subject data listings.

Body weight is measured 3 times at every visit. The median weight will be used in summary tables or figures.

## 3.2. Definitions of Analysis Populations

### 3.2.1. Safety Population

The Safety Population includes all randomized subjects who receive at least 1 dose of study drug (ALKS 3831 or olanzapine). The Safety Population will be used for the safety analyses.

### 3.2.2. Efficacy Population

The Full Analysis Set (FAS) includes all subjects in the Safety Population who have at least one postbaseline weight assessment. The FAS will be used for the efficacy analyses, unless specified otherwise.

The Early Weight Gain population includes all subjects in the FAS population who gain weight (>0 kg) at Week 1.

Stage 1 Subjects are defined as first 200 FAS subjects who have completed or prematurely discontinued before the interim analysis.

Stage 2 Subjects are defined as FAS subjects who completed or prematurely discontinued after the interim analysis.

## 3.3. Disposition

The number and percentage of subjects completing or prematurely discontinuing the study including reasons for discontinuation will be summarized overall and by treatment group for the following.

- Subjects who were randomized
- Subjects in the Safety Population
- Subjects in the FAS Population
- Subjects in the Early Weight Gain Population
- Subjects who completed the double-blind treatment period

- Subjects who discontinued the study during the double-blind treatment period along with reasons for discontinuation

## 3.4. Demographics and Baseline Characteristics

Demographics and baseline characteristics such as sex, age, race, ethnicity, weight, and body mass index (BMI) and psychiatric history will be summarized overall and by treatment group for the Safety Population, the FAS Population, and the Early Weight Gain Population. If there are heterogeneities between study groups in any of the subject characteristics that are of clinical importance or could affect the treatment outcome, the impact of the imbalances will be investigated and, if necessary, appropriate adjustments made in the efficacy and safety analyses.

Medical history will be summarized for the Safety Population by treatment group and overall using the number of observations and percentage of subjects reporting each category.

## 3.5. Protocol Deviation

Subjects with major protocol deviations in the following categories will be summarized by treatment group along with supportive listings for each category.

- Did not meet the inclusion/exclusion criteria
- Received prohibited medications
- Lack of adherence with study medication, as defined by subjects taking less than 70% of the protocol-specified amount of study medication
- Randomization or dosing error

## 3.6. Prior and Concomitant Medication

Prior medications are defined as medications taken prior to the first dose of study drug. Concomitant medications are defined as medications taken on or after the first dose study drug. All medications as documented by the investigator will be coded using the World Health Organization Anatomical Therapeutic Chemical (WHO-ATC) drug dictionary Enhanced Extended with Herbal version WHODRUG_C / 161E+H.

Prior and concomitant medications will be summarized by the preferred drug name for the Safety Population. Concomitant medications that are taken during the double-blind treatment period will be included in the summary table. For subjects who discontinue study drug prematurely and return for monthly visits, all antipsychotic medications that are taken after the last dose of study medication will be summarized separately. For the summary tables, if a subject has taken a prior or concomitant medication more than once, the subject will be counted only once for the medication. All reported medications (including those initiated after the last dose of double-blind study medication) will be included in the listing.

## 3.7. Treatment Adherence Rate and Extent of Exposure

### 3.7.1. Treatment Adherence Rate

Treatment adherence to the daily dosing schedule during the double-blind treatment period will be summarized by treatment group for the Safety Population. Treatment adherence will be calculated as follows:

$$100 \times \frac{\text{Total tablets dispensed} - \text{total tablets returned} - \text{total tablets lost}}{\text{Total tablets scheduled to be taken}}$$

### 3.7.2. Duration of study drug administration

Duration of study drug administration (ALKS 3831 and OLZ) is defined as the number of days from the date of the first dose of study drug taken to the date of the last dose taken, inclusive (ie, last dose date – first dose date + 1). Duration of study drug administration will be summarized for the Safety Population by treatment group.

The overall mean and modal dose of olanzapine will be summarized by treatment group. Number and percentage of subjects will be summarized by their Visit 6 (Week 4) dose level and by final dose level.

## 3.8. Efficacy Analyses

### 3.8.1. General Considerations

All statistical tests will be 2-sided hypothesis tests performed at the 5% significance level. All confidence intervals will be 2-sided 95% confidence intervals.

Baseline for the primary efficacy analysis is defined as the last non-missing observation on or before the first dose of study drug in the double-blind treatment period.

Only the on-treatment weight assessments will be included in the analyses, unless specified otherwise.

### 3.8.2. Primary Efficacy Endpoints

The primary efficacy endpoints are

- percent change from baseline in body weight at Week 24 in FAS population
- proportion of subjects with ≥10% weight gain at Week 24 in FAS population

#### 3.8.2.1. Primary Efficacy Endpoint: Percent Change from Baseline in Body Weight at Week 24 in the FAS population

##### 3.8.2.1.1. Primary Analysis

The primary endpoint will be analyzed by analysis of covariance (ANCOVA) with the multiple imputation (MI) method for handling of missing values. Only the on-treatment measurements will be included in the analysis. The following steps will be performed:

1. The missing data on body weight in Stage 1 subjects will be imputed using MI. If there is any subject whose missing data pattern is non-monotonic (defined as having missing data in between visits), the Markov Chain Monte Carlo (MCMC) method will be used to impute the data to a monotonic missing pattern (i.e., if the data is missing at the current visit, the data at all following visits are missing). A single Markov chain will be used with 200 burn-in iterations, 100 iterations between imputations and a Jeffrey's prior. Next, the missing data will be imputed sequentially by each visit using a regression method. The imputation regression model includes treatment group, race (Black or African American, Non-Black or African American) and baseline age (<30, ≥30 years) as factors, and body weight at all previous visits (including baseline weight) as covariates. At least five hundred imputations will be carried out. More imputations may be performed to improve the precision. The initial seed for both steps of the imputation will be 1104078.
2. Percent change from baseline in weight at Week 24 of each of these multiply imputed datasets will be analyzed by the ANCOVA model with treatment group, race (Black or African American, Non-Black or African American), and baseline age (<30, ≥30 years) as factors, and the baseline weight as covariate.
3. Results from Step 2 will be combined using Rubin's method to get the Stage 1 results, including estimated treatment effect, standard error and test statistic $z_1$.
4. Repeat Steps 1 to 3 for Stage 2 Subjects. Estimated treatment effect, standard error and test statistic $z_2$ will be obtained.
5. To adjust for the IA for sample size re-estimation, the independent test statistic from two stages will be combined using the CHW method with a fixed weight of $\sqrt{0.5}$ based on the original sample size (Cui et al, 1999):

$$Z^*_{chw} = \sqrt{t_1}z_1 + \sqrt{1-t_1}z_2$$

where $t_1 = 200/400$,

In addition, to estimate the treatment effect, the same MI procedure will be performed based on the entire dataset including all FAS subjects. Five hundred imputations will be carried out. The results will be analyzed by the same ANCOVA model and combined using Rubin's method. Estimated treatment effect, standard error (SE), and 95% CIs will be reported.

Descriptive statistics of body weight and percent change from baseline at each visit (including last assessment during treatment period) will be presented by treatment group based on observed and MI-imputed data.

#### 3.8.2.1.2. Sensitivity Analysis

To assess the robustness of the primary analysis, the following sensitivity analyses will be performed:

- To assess the potential impact of missing data due to missing not at random (MNAR), the delta-adjusted Pattern Mixture Model (Ratitch et al, 2013) will be conducted to assess the impact of missing data. It incorporates the clinical assumption that OLZ subjects who discontinue at a given time point would have, on average, their unobserved weight gain decreased by some amount δ compared with the observed weight gain of subjects on the OLZ arm who continue to the next time point. Subjects who discontinue from the ALKS 3831 arm would have the same weight gain trajectory as the ALKS 3831 subjects who stay on the study. A sequential regression-based MI procedure will be used to incorporate the assumption and to allow uncertainty in the imputations to be reflected appropriately in the analysis. The imputation model will include the measurement at the current time point as the response variable, and the measurements at the previous time points, the baseline, race (Black or African American, Non-Black or African American), and age (<30, ≥30 years) as covariates. Five hundred imputations will be carried out. For each of the 500 imputed data sets, the same ANCOVA model as in the primary analysis will be fitted to the percent change from baseline at Week 24 to obtain the treatment effect estimate and standard error. Rubin's rule will be used to combine the treatment effect estimates and standard errors across imputations. The shift parameters will account for 10%, 20%, 30%, 40%, 50%, and up to 100% of the observed treatment difference between OLZ and ALKS 3831 from the primary analysis of the percent change from baseline in body weight at Week 24 in FAS population.
- To further assess the potential impact of missing data due to missing not at random (MNAR), the primary analysis will be repeated including both on-treatment and off-treatment weight assessments after premature discontinuation of study drug.
- The ANCOVA and MI approach may adjust for additional covariates and/or factors, including but not limited to study region or study sites.
- Average treatment difference across visits through Week 24 between ALKS 3831 and OLZ will be analyzed using mixed model with repeated measurements (MMRM) model. The MMRM model will include treatment, visit, treatment-by-visit interaction term, race (Black or African American, Non-Black or African American) and age (<30, ≥30 years) as categorical fixed effects; baseline weight will be included as a covariate. An unstructured covariance structure will be applied. The Kenward-Roger approximation (Kenward and Roger 1997) will be used to adjust the denominator degree of freedom. The analysis will be performed on all observed post-randomization on-treatment weight measurements without imputation of missing data.

#### 3.8.2.1.3. Additional Analyses

To explore the time point (visit) at which the percent change in body weight from baseline has stabilized and is constant during the 24-week double blind treatment period, two methods will be performed.

- Stepwise tests of linear trend (Maganti et al, 2008) will be performed using a linear mixed model (LMM) approach. The LMM model includes treatment, race (Black or African American, Non-Black or African American), and age (<30, ≥30 years) as fixed effects; visit and treatment-by-visit interaction as random effects; and baseline weight as the covariate. In this model, visit will be treated as a continuous variable. The slope of each treatment group will be estimated from the model. Denominator degrees of freedom will be estimated using the Kenward–Roger approximation (Kenward and Roger 1997). The null hypothesis is that there is no linear trend, ie, the slope of the regression line equals zero. The alternative is that there is a linear trend. The first test of the slope uses the entire range of time points. If the slope is significantly different from zero, the second test excludes the earliest time point. The testing continues, successively excluding the next earliest time point from the start of the study, until the slope is no longer significantly different from zero or until only three time points remain in the model. If the final test is not statistically significant, then the first time point included in that test is considered to be the time point over which a constant weight is attained. If the final test includes three time points and is still statistically significant, then no stable weight is considered to have been attained by the end of the study.
- Helmert contrast (Maganti et al, 2008) will be performed using a similar MMRM approach as outlined in Section 3.8.2.1.2. The first contrast compares the mean change in body weight at the first time point to the pooled mean change in body weight over all remaining time points. The second contrast compares the mean at the second time point to the pooled mean over all remaining time points. Testing continues until the contrast is not statistically significant. The first time point included in this last contrast is concluded to be the time interval on which a stable weight is attained. If the final test is still statistically significant, then no stable weight is considered to have been attained by the end of the study.

#### 3.8.2.2. Primary Efficacy Endpoint: Proportion of Subjects with ≥10% Weight Gain at Week 24 in the FAS Population

##### 3.8.2.2.1. Primary Analysis

The primary analyses will be carried out using a logistic regression model based on MI for missing data as described in Section 3.8.2.1.1 . Proportion of subjects with ≥10% Weight Gain at Week 24 will be derived based on the same complete datasets obtained from the multiple imputation procedure. The logistic regression model will include the treatment group, race (Black or African American, Non-Black or African American), and age (<30, ≥30 years) as factors; and the baseline weight as the covariate. For subjects who prematurely discontinue the study drug but are willing to come back for weight assessments, only on-treatment weight assessments will be included in the primary analyses.

Similarly to the primary efficacy endpoint of percent change from baseline body weight at Week 24, the independent test statistic before and after the interim analysis will be combined using the CHW method with a fixed weight of $\sqrt{0.5}$ based on the original sample size. The treatment effects and 95% CIs will be reported as described in Section 3.8.2.1.1.

#### 3.8.2.2.2. Sensitivity analysis

To assess the robustness of the primary analysis of the primary endpoint, the following sensitivity analyses will be performed:

- To assess the potential impact of missing data due to missing not at random (MNAR), the delta-adjusted Pattern Mixture Model in Section 3.8.2.1.2 will be conducted to assess the impact of missing data. Proportion of subjects with ≥10% Weight Gain at Week 24 will be derived based on the same complete datasets obtained from the multiple imputation procedure in Section 3.8.2.1.2. The logistic regression model as in the primary analysis will be fitted to obtain the treatment effect estimate and standard error. Rubin's rule will be used to combine the treatment effect estimates and standard errors across imputations.
- To further assess the potential impact of missing data due to missing not at random (MNAR), the approach described in Section 3.8.2.2.1 will be repeated including both on-treatment and off-treatment weight assessments after premature discontinuation of study drug.
- The logistic regression model may also adjust for additional covariates and/or factors, including but not limited to, study region or study sites.
- Average treatment difference across visits through Week 24 between ALKS 3831 and OLZ will be analyzed.

#### 3.8.2.2.3. Additional Analyses

To assess the impact of different cutoffs for weight gain (ie, 5%, 7%, 10%, etc), a plot of cumulative responder distribution at Week 24 for each treatment group will be presented. The horizontal axis will present percent weight gain, and the vertical axis will present the proportion of subjects with weight gain. The analysis will be based on MI imputation for missing data.

### 3.8.3. Key Secondary Endpoint

The key secondary endpoint is the proportion of subjects with ≥7% weight gain at Week 24. The primary analysis will be carried out using a logistic regression model based on MI without CHW adjustment as described in Section 3.8.2.2.1. The sensitivity analysis will be the same as Section 3.8.2.2.2.

### 3.8.4. Multiple Comparison/ Multiplicity

The study will be claimed positive when the primary analyses of both co-primary efficacy endpoints are statistically significant based on the CHW method to adjust for the unblinded interim analysis for sample size re-estimation; therefore, both co-primary endpoints will be tested at a full α level of 0.05. A fixed sequence approach will be used to control overall Type I error rate for the primary and key secondary endpoints. Only if both primary endpoints are statistically significant based on the CHW method will the key secondary endpoint be further tested at an alpha level of 0.05.

### 3.8.5. Examination of Subgroups

Subgroup analyses of the primary endpoints will be performed for each of the following categories:

- Sex (male, female)
- Age (<30 years, ≥30 years)
- Race (Black or African American, Non-Black or African American)
- BMI (<27 kg/m$^2$, ≥27 kg/m$^2$)

The forest plot of the least squares mean difference along with 95% CIs in percent change in body weight from baseline at Week 24 between ALKS 3831 and OLZ will be provided by the subgroup factors listed above.

### 3.8.6. Other Efficacy Endpoints

- Percent change from baseline in body weight at Week 24 in the Early Weight Gain population
- Proportion of subjects with ≥10% weight gain at Week 24 in the Early Weight Gain population
- Change from baseline in fasting lipids (fasting triglycerides, LDL, HDL, total cholesterol), fasting glucose, fasting insulin and HbA1c by visit
- Absolute change in body weight by visit
- Change from baseline in waist circumference by visit
- Change from baseline in PANSS total score and subscales by visit
- Change from baseline in Clinical Global Impression-Severity (CGI-S) by visit
- CGI-I score by visit
- Change from baseline in Impact of Weight on Quality of Life (IWQOL)-Lite total score and subscales by visit
- Change from baseline in EuroQol five dimensions questionnaire (EQ-5D)-5L score by visit

The following endpoints will be analyzed similarly by ANCOVA model and MI for missing data as noted in Section 3.8.2.1.1, without CHW adjustment:

- Percent change from baseline in body weight at Week 24 in the Early Weight Gain population
- Absolute change in body weight by visit
- Change from baseline in waist circumference by visit

Proportion of subjects with ≥10% weight gain at Week 24 in the Early Weight Gain population will be analyzed similarly by logistic regression model and MI for missing data as described in Section 3.8.2.2.1, without CHW adjustment.

Change from baseline in fasting lipids (fasting triglycerides, LDL, HDL, total cholesterol), fasting glucose, fasting insulin and HbA1c will be analyzed by the same MMRM model as described in Section 3.8.2.1.2. The model will include treatment, visit, treatment-by-visit interaction term, race (Black or African American, Non-Black or African American) and age (<30, ≥30 years) as categorical fixed effects; the baseline value will be included as the covariate.

MMRM model will also be applied to the following endpoints:

- Change from baseline in PANSS total score and subscales by visit
- Change from baseline in Clinical Global Impression-Severity (CGI-S) by visit
- Change from baseline in Impact of Weight on Quality of Life (IWQOL)-Lite total score and subscales by visit
- Change from baseline in EuroQol five dimensions questionnaire (EQ-5D)-5L score by visit

The model will include treatment, visit, treatment-by-visit interaction term as categorical fixed effects; the baseline value will be included as a covariate.

PANSS responder status (<20%, ≥20% improvement over baseline PANSS), CGI-I score, and CGI-I responder status (≤2 versus >2) at each visit will be summarized by descriptive statistics.

IWQOL-Lite scale is a 31-item self-report measure of obesity-specific quality of life. The IWQOL-Lite provides an overall total score as well as scores on five domains: (1) physical function, (2) self-esteem, (3) sexual life, (4) public distress, and (5) work. The raw IWQOL scores will be transformed as follows (Tessier et al, 2012) before analyzed by the MMRM method. The transformed scores range from 0 to 100, with 100 representing the best and 0 the most impaired quality of life.

$$\text{Transformed Score} = \frac{\text{maximum theoretical score} - \text{actual score}}{\text{test score range}} \times 100$$

The EQ-5D-5L is a validated quality of life questionnaire developed by the EuroQol Group in order to provide a simple, generic utility measure for characterizing current health states of patients. EQ-5D-5L is designed for self-completion by subjects. It consists of 2 parts – the EQ-5D-5L descriptive system and the visual analogue scale (VAS).

The EQ-5D-5L descriptive system comprises 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. Each dimension has 5 levels.

The 5 dimensional 5-level systems are converted into an index value. Values for theoretically possible health states are calculated using a regression model and weighted according to the social preferences of the U.S. population (Van Hout et al, 2012). The VAS records the respondent's self-rated health on a vertical visual analogue scale. The VAS 'thermometer' has 100 (Best imaginable health state) at the top and 0 (Worst imaginable health state) at the bottom. This information can be used as a quantitative measure of health outcomes as judged by the individual respondents. EQ-5D-5L self-reported VAS data generates information on the self-perceived overall health-related quality of life.

Additional analyses of metabolic laboratory parameters, body weight, and PANSS total score are included in the Exploratory Statistical Analysis Plan (ESAP) conducted outside the scope of this SAP.

### 3.8.7. Exploratory Endpoints

Body composition is assessed in the substudy at Visit 2 (Week 1) and Visit 17/ ET (Week 24) for a subset of subjects by the following exploratory endpoints:

- Absolute change from baseline in lean mass
- Absolute change from baseline in fat mass
- Change from baseline in percent body fat

These endpoints will be analyzed by the ANCOVA method adjusting for race (Black or African American, Non-Black or African American), and baseline age (<30, ≥30 years) as factors and corresponding baseline value as covariate for the subset of subjects with baseline and post-baseline body composition assessments.

## 3.9. Safety Analysis

### 3.9.1. Adverse Events

Adverse events will be coded by System Organ Class and Preferred Term using the Medical Dictionary for Regulatory Activities (MedDRA®) Version 21.0. The verbatim term will be included in the AE listings.

An AE (classified by preferred term) will be considered as treatment-emergent AE (TEAE) if the event is newly occurring or worsening on or after the date of first dose of study drug.

The number and percentage of subjects reporting TEAEs during the double-blind treatment period in each treatment group will be presented by treatment group for the following categories:

- System organ class and preferred term
- Preferred terms in decreasing frequency, and also including the following subsets:
  - Experienced by ≥2% of subjects in any treatment group
  - Experienced by ≥5% of subjects in any treatment group and ≥2 times of OLZ group
- System organ class, preferred term, and severity
- System organ class, preferred term for severe TEAEs
- System organ class, preferred term, and relationship
- System organ class, preferred term for drug-related TEAEs

The number and percentage of subjects reporting TEAEs during the double-blind treatment period between Visit 6 (Week 4) to Visit 17 (Week 24) will be presented by treatment group, OLZ dose levels at Visit 6 and preferred terms.

If more than one AE is coded to the same preferred term for the same subject for the same period, the subject will be counted only once for that preferred term using the most severe and most related occurrence for the summarization by severity and by relationship to study drug.

In addition, the number and percentage of subjects reporting AEs during the Safety Follow-up Period will be tabulated by the system organ class, preferred term and treatment group.

For the subjects who prematurely discontinue the study drug and return for the monthly visits, all AEs collected after discontinuation of the study drug will be summarized by system organ class, preferred term and treatment group.

The number and percentage of subjects who have serious adverse events (SAE) and AEs leading to premature discontinuation from the treatment will be summarized by preferred term and treatment group and will be sorted by decreasing frequency for ALKS 3831 group.

#### 3.9.1.1. Other significant AEs

In addition, incidence of a selected subset of relevant AEs in this class of drugs (eg, Extrapyramidal symptom (EPS) TEAEs, AEs associated with abuse potential, AEs associated with drug withdrawal, and suicide related events, etc.) will be summarized by treatment group and preferred term. The selection of AEs per subset will be based on the preferred term, Standardized MedDRA queries (SMQs) or customized MedDRA queries (CMQs).

### 3.9.2. Clinical Laboratory Parameters

Laboratory parameters will be presented in conventional (ie, US) units. Unless specified otherwise, only scheduled laboratory parameters will be included in the summaries. All laboratory data, including those collected at unscheduled visits, will be included in the listings.

Laboratory results (baseline and change from baseline) for the Safety Population during double-blind treatment period for chemistry and hematology parameters will be summarized by treatment group and by visit.

Number (percentage) of subjects with potentially clinically significant (PCS) values at any post-baseline visit, will be summarized by treatment group. PCS criteria are presented in Table 1. The denominator is all subjects with non-PCS baseline and at least one post-baseline assessment in the Safety Population and the numerator is the number of subjects with non-PCS baseline and PCS at post-baseline. All PCS values including baseline PCS values will be included in supportive listings.

Shift tables for metabolic parameters (fasting glucose, total cholesterol, LDL, HDL, Triglycerides, and HbA1c) and liver function tests will be presented. The criteria are summarized in Table 2, Table 3 and Table 4.

**Table 1: Criteria of Potentially Clinically Significant (PCS) Abnormality for Selected Analytes**

| Parameters | Criteria |
|---|---|
| **Chemistry** | |
| Albumin | <2.5 g/dL |
| Alkaline Phosphatase (U/L) | ≥3 × ULN |
| Alanine Aminotransferase (U/L) | ≥3 × ULN |
| Aspartate Aminotransferase (U/L) | ≥3 × ULN |
| Bilirubin, Total | ≥2.0 mg/dL |
| Blood Urea Nitrogen | >30 mg/dL |
| Cholesterol, Random | >300 mg/dL |
| Cholesterol, Fasting | ≥240 mg/dL |
| Cholesterol, HDL Fasting | ≤30 mg/dL |
| Cholesterol, LDL Fasting | ≥160 mg/dL |
| Creatine Kinase (U/L) | ≥3 × ULN |
| Creatinine | ≥2.0 mg/dL |
| Glucose, Random | <50 mg/dL or<br>≥200 mg/dL |
| Glucose, Fasting | <50 mg/dL or<br>≥126 mg/dL |
| Potassium | <3 mmol/L or<br>>5.5 mmol/L |
| Lactate Dehydrogenase (U/L) | >3 × ULN |
| Prolactin (Female) | >30 ng/mL |
| Prolactin (Male) | >20 ng/mL |
| Sodium | <130 mmol/L or<br>>150 mmol/L |
| Triglycerides, Fasting (Female) | ≥120 mg/dL |
| Triglycerides, Fasting (Male) | ≥160 mg/dL |

**Table 1: Criteria of Potentially Clinically Significant (PCS) Abnormality for Selected Analytes (Continued)**

| Parameters | Criteria |
|---|---|
| **Hematology** | |
| Eosinophils | $>1.0 \times 10^3/\mu L$ |
| Hematocrit (Female) | ≤32% |
| Hematocrit (Male) | ≤37% |
| Neutrophils, Absolute | $<1.5 \times 10^3/\mu L$ |
| Platelets | $<75.0 \times 10^3$ cells/μL or<br>$\geq 700.0 \times 10^3$ cells/μL |
| Leukocytes | $\leq 2.8 \times 10^3/\mu L$ or<br>$\geq 16.0 \times 10^3/\mu L$ |

**Table 2: Shifts Category from Baseline to Any Post-Baseline for Selected Lipid Parameters**

| |
|---|
| **Total cholesterol (fasting) mg/dL** |
| Normal (<200) to high (≥240) |
| Borderline (≥200 and < 240) to high (≥240) |
| Normal/Borderline (<240) to high (≥240) |
| Normal (<200) to borderline/high (≥200) |
| Increase ≥40 mg/dL |
| **LDL Cholesterol (fasting) mg/dL** |
| Normal (<100) to high (≥160) |
| Borderline (≥100 and <160) to high (≥160) |
| Normal/borderline (<160) to high (≥160) |
| Normal (<100) to borderline/high (≥100) |
| Increase ≥30 mg/dL |
| **HDL Cholesterol (fasting) mg/dL** |
| Normal (≥40) to low (<40) |
| Decrease ≥20 mg/dL |
| **Triglycerides (fasting) mg/dL** |
| Normal (<150) to high (≥200) |
| Normal (<150) to very high (≥500) |
| Borderline (≥150 and <200) to high (≥200) |

**Table 2: Shifts Category from Baseline to Any Post-Baseline for Selected Lipid Parameters (Continued)**

| **Triglycerides (fasting) mg/dL** |
|---|
| Borderline (≥150 and <200) to very high (≥500) |
| Normal/borderline (<200) to high (≥200) |
| Normal/borderline (<200) to very high (≥500) |
| Normal (<150) to borderline/high/very high (≥150) |
| Increase ≥50 mg/dL |

**Table 3: Shift Category from Baseline to Any Post-baseline in Glucose and HbA1c**

| **Serum glucose (fasting) mg/dL** |
|---|
| Normal (<100) to high (≥126) |
| Impaired (≥100 and <126) to high (≥126) |
| Normal/Impaired (<126) to high (≥126) |
| Increase ≥10 mg/dL |
| **HbA1c%** |
| Shift from baseline (<5.7%) to post-baseline ≥5.7% |
| Shift from baseline (<5.7%) to post-baseline ≥5.7% and <6.5% |
| Shift from baseline (<5.7%) to post-baseline ≥6.5% |

**Table 4: Shift Category from Baseline to Any Post-Baseline in Liver Function Test**

| **Alanine Aminotransferase (ALT) (U/L)** |
|---|
| Shift from Normal to ≥3 × ULN |
| Shift from Normal to ≥5 × ULN |
| Shift from Normal to ≥10 × ULN |
| **Aspartate Aminotransferase (AST) (U/L)** |
| Shift from Normal to ≥3 × ULN |
| Shift from Normal to ≥5 × ULN |
| Shift from Normal to ≥10 × ULN |
| **Bilirubin, Total (mg/dL)** |
| Shift from Normal to >1 × ULN |
| Shift from Normal to ≥2 × ULN |

### 3.9.3. Vital signs, and Electrocardiograms

#### 3.9.3.1. Vital Signs

Descriptive statistics for vital signs and changes from baseline values at each scheduled time point will be presented by treatment group for the double-blind period.

Vital sign values will be considered PCS if they meet both the observed-value criteria and the change from-baseline criteria listed in Table 5. The number and percentage of subjects with PCS post-baseline values will be tabulated by treatment group for the double-blind treatment period. The percentages will be calculated relative to the number of subjects with available non-PCS baseline values and at least 1 post-baseline assessment for the double-blind period. The numerator will be the total number of subjects with available baseline values and at least 1 PCS post-baseline value for double-blind period. A supportive tabular display of subjects with PCS post-baseline values will be provided, including the subject ID number, and baseline and all post-baseline (including non-PCS) values.

All vital signs will be presented in the subject data listing.

Orthostatic hypotension (20/10) is defined as a fall in systolic blood pressure of at least 20 mmHg and a fall in the diastolic blood pressure of at least 10 mmHg upon standing from supine. Orthostatic hypotension (30) is defined as a fall in systolic blood pressure of at least 30 mmHg upon standing from supine.

Orthostatic tachycardia is defined as a heart rate increase of 30 beats per minute (bpm) or more upon standing from supine, or over 120 bpm upon standing.

The number and percentage of subjects with orthostatic hypotension or orthostatic tachycardia occurring at any post-baseline visit will also be summarized by treatment group for the double-blind period.

**Table 5: Criteria for Potentially Clinically Significant (PCS) Blood Pressure or Pulse Rate**

| Parameter | Criteria |
|---|---|
| Supine Systolic Blood Pressure | ≤90 and decrease ≥20 mm Hg<br>≥180 and increase ≥20 mm Hg |
| Supine Diastolic Blood Pressure | ≤50 and decrease ≥15 mm Hg<br>≥105 and increase ≥15 mm Hg |
| Supine Heart Rate | ≤50 and decrease ≥15 bpm<br>≥120 and increase ≥15 bpm |

#### 3.9.3.2. Electrocardiograms

Descriptive statistics for ECG parameters (heart rate, PR interval, RR interval, QRS interval, QT interval, and QTc interval) at baseline and change from baseline values at each assessment time point and at the end of the double-blind treatment period will be presented by treatment group for the double-blind period. QTc interval will be calculated using both Bazett

(QTcB = QT/(RR)$^{1/2}$) and Fridericia (QTcF = QT/(RR)$^{1/3}$) corrections; if RR is not available, it will be replaced with 60/HR in the correction formula.

Electrocardiogram parameter values are considered PCS if they meet or exceed the higher-limit PCS criteria listed in Table 6. The number and percentage of subjects with PCS post-baseline ECG values will be tabulated by treatment group for the double-blind treatment period. The percentages will be calculated relative to the number of subjects with non-PCS baseline values and at least 1 post-baseline assessment for the double-blind treatment period. The numerator is the total number of subjects with non-PCS baseline values and at least 1 PCS post-baseline value for the double-blind treatment period. A supportive tabular display of subjects with PCS post-baseline values will be provided, including the subject ID number, study center number, the baseline value, all post-baseline (including non-PCS) values, and change from baseline.

**Table 6: Criteria for Potentially Clinically Significant (PCS) QTcB and QTcF**

| Parameter | Criteria |
|---|---|
| QTcB and QTcF | >450 to ≤480 msec |
| | >480 to ≤500 msec |
| | >500 msec |
| | Change from baseline >30 to ≤60 msec |
| | Change from baseline >60 msec |

### 3.9.4. Abnormal Movement Scales

Extra pyramidal symptoms (EPS) will be evaluated as AEs and also as assessed by abnormal movement scales. Abnormal movement scales will include the following: Abnormal Involuntary Movement Scale (AIMS), Barnes Akathisia Rating Scale (BARS), and Simpson-Angus Scale (SAS).

For all abnormal movement scales, total scores and subscale scores will be summarized by treatment group at each visit for the absolute value and for changes from baseline.

Number and percentage of subjects meeting the criteria for treatment-emergent Parkinsonism (SAS total score>3), for treatment-emergent akathisia (BARS global clinical assessment of akathisia score ≥2), for treatment-emergent dyskinesia (AIMS score ≥3 on any of the first 7 items, or a score ≥2 on two or more of any of the first 7 items) at any post-baseline visit will be summarized by treatment group.

A listing will be provided for every abnormal movement scale. Listing for treatment-emergent EPS will be provided.

### 3.9.5. Columbia-Suicide Severity Rating Scale (C-SSRS)

The C-SSRS is a questionnaire used to measure the presence and intensity of suicidal ideation and behavior.

Suicidal behavior and suicidal ideation will be summarized descriptively. The number of subjects with suicidal ideation and suicidal behavior during the double-blind treatment period will be summarized by treatment group when applicable.

Supportive tabular display of subjects with all values will be provided, including subject ID number, treatment group, visit number, intensity of suicidal ideation, suicidal behavior type, and lethality of suicidal behavior.

**Table 7: C-SSRS Categories for Analysis**

| Category | C-SSRS Item response is "YES" |
|---|---|
| Suicidal behavior | Preparatory acts or behavior<br>Aborted attempt<br>Interrupted attempt<br>Actual attempt<br>Complete suicide |
| Suicidal ideation | Wish to be dead<br>Non-specific active suicidal thoughts<br>Active suicidal ideation with any methods (not plan) without intent to act<br>Active suicidal ideation with some intent to act, without specific plan<br>Active suicidal ideation with specific plan and intent |
| Non-suicidal self-injurious behavior | Non-suicidal self-injurious behavior |

## 3.10. Pharmacokinetic/ Pharmacodynamic Data Analysis

Subject listings for the PK sampling time and concentrations of olanzapine, samidorphan, and metabolites of interest will be provided. PK data obtained from plasma samples collected in this study may be included in a subsequent population PK analysis or other post-hoc analyses conducted outside the scope of this SAP.

# 4. INTERIM ANALYSIS

One unblinded interim analysis will be conducted by an independent statistical center (ISC) when 50% of subjects (N=200) have completed the double-blind treatment period or have discontinued from the study. The details regarding the conduct of the unblinded interim analysis will be specified in the interim analysis operational plan and the statistical testing methodology will be specified in the interim analysis plan (IAP).

The conditional power (CP) will be computed for both primary endpoints. Based on the conditional power, it will be assigned into one of 3 zones: favorable (CP ≥90%), promising (30% ≤ CP <90%), or unfavorable (CP< 30%). The target sample size of 200 per arm will be increased to 270 subjects per arm when the conditional power for at least one primary endpoint is in the promising zone and neither of the primary endpoints are in the unfavorable zone. Otherwise, the target sample size will be maintained at 200 subjects per arm. Overall Type 1 error for the final analysis are maintained using CHW method (Cui et al, 1999) by combining the Z test statistics from two indpendent stages (before and after the interim analysis) with a fixed weight based on the original sample size. The weighted CHW statistics always controls type I error, as it does not depend on the design specifications, such as the sample size adaptation algorithm, the maximum

number to be increased, and the final sample size (Hung et al, 2014).There will be no early stopping for efficacy or futility. The sponsor will not be informed of any unblinded interim analysis results except for a recommendation to either increase the sample size to 540 subjects (270 per arm) or to leave the sample size unchanged.

# 5. CHANGES IN CONDUCT OR PLANNED ANALYSES FROM THE PROTOCOL

There are no major deviations from the protocol-specified analysis.

# 6. DEFINITIONS AND CONVENTIONS FOR HANDLING OF THE DATA

Dataset specifications will be provided in a separate document.

## 6.1. Analysis Visit Windows

Scheduled analysis visits are visits upon scheduled time points as specified in the protocol (Table 3 Schedule of Visits and Assessments).

Scheduled analysis visits during the study period will be the same as the nominal visits collected in the eCRF. There will be one valid value of assessment kept for each scheduled analysis visit in summary/ analysis statistics.

Unscheduled visits are visits with data collected outside of the scheduled time point. Unscheduled visits will not be used for by-visit summary/analysis statistics unless specified otherwise.

All unscheduled visit data as collected in eCRFs will be included in listings.

Visit Day is calculated as (visit date) – (date of the first dose of study drug) + 1.

Last post-baseline values are defined as the last valid post-baseline values collected for each subject during the double-blind treatment period.

Any post-baseline values are defined as any valid values collected from scheduled visits if applicable.

## 6.2. Handling of Partial Dates of Concomitant Medication

Partial start dates of prior and concomitant medications will be assumed to be the earliest possible date consistent with the partial date. Partial stop dates of prior and concomitant medications will be assumed to be the latest possible date consistent with the partial date. In the case of a completely missing stop date, medication will be assumed to be ongoing.

## 6.3. Handling of Safety Data

All efforts should be made to obtain any missing information from the investigator. For C-SSRS, vital signs, laboratory testing (chemistry, hematology, urinalysis), and 12-lead ECGs, only observed data will be used for analyses, and missing data will not be imputed.

# 7. GENERAL STATISTICAL METHODOLOGY

In general, summary statistics (n, mean, standard deviation [SD], median, minimum, and maximum for continuous variables, and number and percentage of subjects in each category for categorical variables) will be provided by treatment group. All summary tables will be based on observed data, and missing values will not be imputed unless otherwise indicated. Measurements collected from unscheduled visits or repeated assessments will not be included in the by-visit summary tables or figures, the analyses for the PCS post-baseline values, and subject listings. Source data for the summary tables and statistical analyses will be presented as subject data listings.

## 7.1. Reporting Precision

Summary statistics will be presented to the following degree of precision, unless otherwise specified:

**Table 8: Degree of Precision**

| Statistics | Degree of Precision |
|---|---|
| Mean, Median, Confidence limit boundaries | One more than the raw data, up to 3 decimal places |
| Standard deviation, Standard error | One more than the mean, up to 3 decimal places |
| Minimum, Maximum | The same as the raw data, up to 2 decimal places |
| p-value | Rounded to 3 decimal places and therefore presented as 0.xxx; p-values smaller than 0.001 as '<0.001'; p-values greater than 0.999 as '>0.999' |
| Percentage | One decimal place. A percentage of 100% will be reported as 100%. Percentages of zero will be reported as 0 |

Fractional numeric values will be presented with a zero to the left of the decimal point (for example, 0.12 – 0.30).

For weight, height, and BMI, one decimal place will be used for summary statistics.

# 8. PROGRAMMING SPECIFICATIONS

Programming specifications will be provided in a separate document.

# 9. MOCK TABLES, LISTINGS AND GRAPHS (TLGS)

Mock-up tables and listings will be provided in a separate document.

# 10. REFERENCES

Alkermes ALK3831-A303 Study Protocol Amendment 5.0 Date: 18 Sep 2018

Alkermes ALK3831-A303 Interim Analysis Plan Final Version 2.0 Date: 05 May 2017

Alkermes ALK3831-A303 Interim Operational Plan Final Version 2.0 Date: 05 May 2017

Alkermes ALK3831-A303 Exploratory Statistical Analysis Plan Final Version 1.0 Date: 18 Sep 2018

Cui L, Hung H, Wang S. Modification of sample size in group sequential clinical trials. *Biometrics.* 1999;55:853-857.

Hung HM, Wang SJ, Yang P. Some challenges with statistical inference in adaptive designs. *J Biopharm Stat.* 2014;24(5):1059-1072. doi: 10.1080/10543406.2014.925911.

Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics.* 1997;53(3):983-997.

Maganti L, Panebianco DL, Maes AL. Evaluation of methods for estimating time to steady state with examples from phase 1 studies. *AAPS J.* 2008;10(1):141-147.

Ratitch B, O'Kelly M, Tosiello R. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharm Stat.* 2013;12(6):337-347. doi: 10.1002/pst.1549.

Tessier A, Zavorsky GS, Kim dJ, Carli F, Christou N, Mayo NE. Understanding the Determinants of Weight-Related Quality of Life among Bariatric Surgery Candidates. *J Obes.* 2012;2012:713426. doi: 10.1155/2012/713426.

van Hout B, Janssen MF, Feng YS, et al. Interim scoring for the EQ-5D-5L: mapping the EQ-5D-5L to EQ-5D-3L value sets. *Value Health.* 2012;15(5):708-715. doi: 10.1016/j.jval.2012.02.008.