Otsuka Pharmaceutical
Development & Commercialization, Inc.

Investigational Medicinal Product

Brexipiprazole

Protocol 331-201-00079
IND No. 103,958

A Phase 3, Multicenter, Randomized, Double-blind, Placebo-controlled Trial to Evaluate the Efficacy, Safety, and Tolerability of Brexpiprazole as Adjunctive Therapy in the Maintenance Treatment of Adults with Major Depressive Disorder

Statistical Analysis Plan

Version 2.0
(Prior to 1st Interim Analysis)

Date: May 13, 2021

Table of Contents

[redacted]

List of In-text Figures

1 Introduction

This statistical analysis plan (SAP) documents the statistical methodology and data analysis algorithms and conventions to be applied for statistical analysis and reporting of efficacy and safety data of Trial 331-201-00079. All amendments to the protocol are taken into consideration in developing this SAP.

2 Trial Objectives

The primary objective of this trial is to compare the efficacy of brexpiprazole (2 to 3 mg/day) to placebo as adjunctive therapy to antidepressant therapy (ADT) for the maintenance treatment in adults with major depressive disorder (MDD). The secondary objective of this trial is to evaluate the safety and tolerability of brexpiprazole (2 to 3 mg/day) as adjunctive therapy to ADT in the proposed subject population with MDD.

3 Trial Design

This is a randomized, double-blind, placebo-controlled trial consisting of a screening phase and three treatment phases. The trial design Schematic is shown in Figure 3-1.

3.1 Screening Phase

Eligibility will be determined during screening and washout of prohibited medications will begin, if applicable. The screening period will be 3 to 42 days. After the screening procedures are completed, eligible subjects will enter Phase A, as appropriate based on the subject's current antipsychotic treatment(s) and required washout of prohibited medication(s).

3.2 Conversion/Prohibited Medication Washout (Phase A)

At Phase A baseline, all subjects must have either 2 or 3 documented inadequate responses to ADT in total for the current episode as defined by the Massachusetts General Hospital Antidepressant Treatment Response Questionnaire (ATRQ). Subjects will be enrolled into the 6- to 8-week, single-blind, Acute Treatment Phase (Phase A) if they meet the entrance criteria. The purpose of the Acute Treatment Phase is to identify subjects who respond to adjunctive brexpiprazole + protocol-specified ADT.

Treatment will consist of single-blind brexpiprazole plus continuation of the ADT taken during the screening period.

Subjects will attend weekly visits during the Acute Treatment Phase (Phase A) and will be evaluated at Weeks 6 to 8 to determine whether they meet response criteria, defined as:

- An improvement on the Montgomery Asberg Depression Rating Scale (MADRS) total score of ≥ 50% from Phase A baseline and
- A Clinical Global Impression - Severity of Illness (CGI-S) score ≤ 3.

Subjects not meeting the response criteria by the Week 8 visit will be discontinued from the trial. Note that all the criteria mentioned for Phase A (in this section), Phase B (Section 3.3), and C (Section 3.4) are blinded to the investigators.

3.3 Single-blind Stabilization (Phase B)

The following subjects will enter Phase B and receive single-blind brexpiprazole + open-label ADT for 12 weeks:

- Subjects who successfully complete Phase A
- Subjects who meet response criteria at the end of Phase A.

During Phase B, investigators will attempt to stabilize subjects on single-blind brexpiprazole for 12 weeks. Visits will occur weekly for; subjects will be assessed for stability at scheduled visits using the following criteria:

- Subjects must show ≥ 50% improvement on the MADRS total score from Phase A baseline and a CGI-S score ≤ 3 for 12 consecutive weeks.
- The dose of brexpiprazole + ADT must be stable for at least the last 4 weeks of Phase B.

Subjects are allowed a maximum of 3 excursions from meeting the stabilization criteria during the 12-week period, as long as the excursions do not occur on 3 consecutive visits or on the last visit of Phase B.

Subjects not meeting stability criteria will be discontinued from the trial.

3.4 Double-blind Maintenance (Phase C)

Subjects eligible for the Double-blind Randomized Withdrawal Phase (Phase C) will be randomized 1:1 to double-blind treatment with brexpiprazole + ADT or placebo + ADT for up to 26 weeks. During Phase C, subjects will be evaluated in the clinic at biweekly visits for the first 2 visits and monthly visits thereafter.

Time-to-relapse (primary efficacy endpoint) is defined as meeting any of the following criteria measured from randomization into Phase C:

- At the same visit, an increase on the MADRS total score of 50% from randomization and CGI-S score ≥ 4, **OR**
- Hospitalization for depression, **OR**

- Discontinuation for lack of efficacy or worsening of depression, **OR**
- Active suicidality, defined as a score of ≥ 4 on the MADRS item 10 **OR** an answer of "yes" on question 4 or 5 on the Columbia-Suicide Severity Rating Scale (C-SSRS) **OR** an answer of yes to any of the questions on the Suicidal Behavior section of the C-SSRS.

Subjects meeting the criteria for relapse will be withdrawn due to lack of efficacy. In addition, functional relapse is defined as a 30% increase in Sheehan Disability Scale (SDS) mean total score from Phase C baseline and at least one SDS sub-score at 4 or greater and a SDS total score great than or equal to 7 when all 3 sub-scores are available (or SDS total score greater than or equal to 5 when work/school item does not apply) at a Phase C visit. In the event that the SDS phase C baseline is 0, the evaluation will be based on the SDS total score and SDS sub-score. Time-to-functional relapse will be the earliest time in Phase C a subject met the above criteria. If a visit is needed between scheduled visits in Phase C, it will be considered an unscheduled visit unless the subject meets criteria for relapse in which case the visit will be considered the end of trial visit and appropriate evaluations for Week 26/End of Phase C/Relapse should occur.

3.5 Follow-up

Subjects who complete the 26-week double-blind treatment phase and those who are discontinued or withdrawn from any trial phase will be prescribed appropriate antipsychotic treatment as per the investigator or the subject's psychiatrist or primary care physician and will be followed up for safety 21 (+ 2) days after the last dose of investigational medicinal product (IMP) via telephone contact or clinic visit at the investigator's discretion.

3.6 Trial Population

The trial population will include male and female outpatients between 18 and 65 years of age at the time of consent, inclusive, with a Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) diagnosis of recurrent MDD and a current major depressive episode confirmed by both the Mini International Neuropsychiatric Interview (MINI) and an adequate clinical psychiatric evaluation. The current episode must be at least 8 weeks in duration.

Subjects must meet all eligibility criteria (inclusion and exclusion criteria) specified in the protocol.

The sponsor reserves the right to utilize external quality oversight methods to ensure the validity of diagnosis, severity of illness, and other factors determining appropriateness of subject selection.

It is anticipated that approximately 1450 subjects will be enrolled into Phase A in order to enroll approximately 700 in Phase B and randomize 450 subjects in Phase C

3.7 Interim Analysis Review Committee

The Interim Analysis Review committee (IARC) is formed to independently review unblinded efficacy data for the planned interim analyses (at approximately 50% and 75% of protocol event accrual) and make recommendations to the sponsor. Details are described in the IARC Charter for the Protocol.

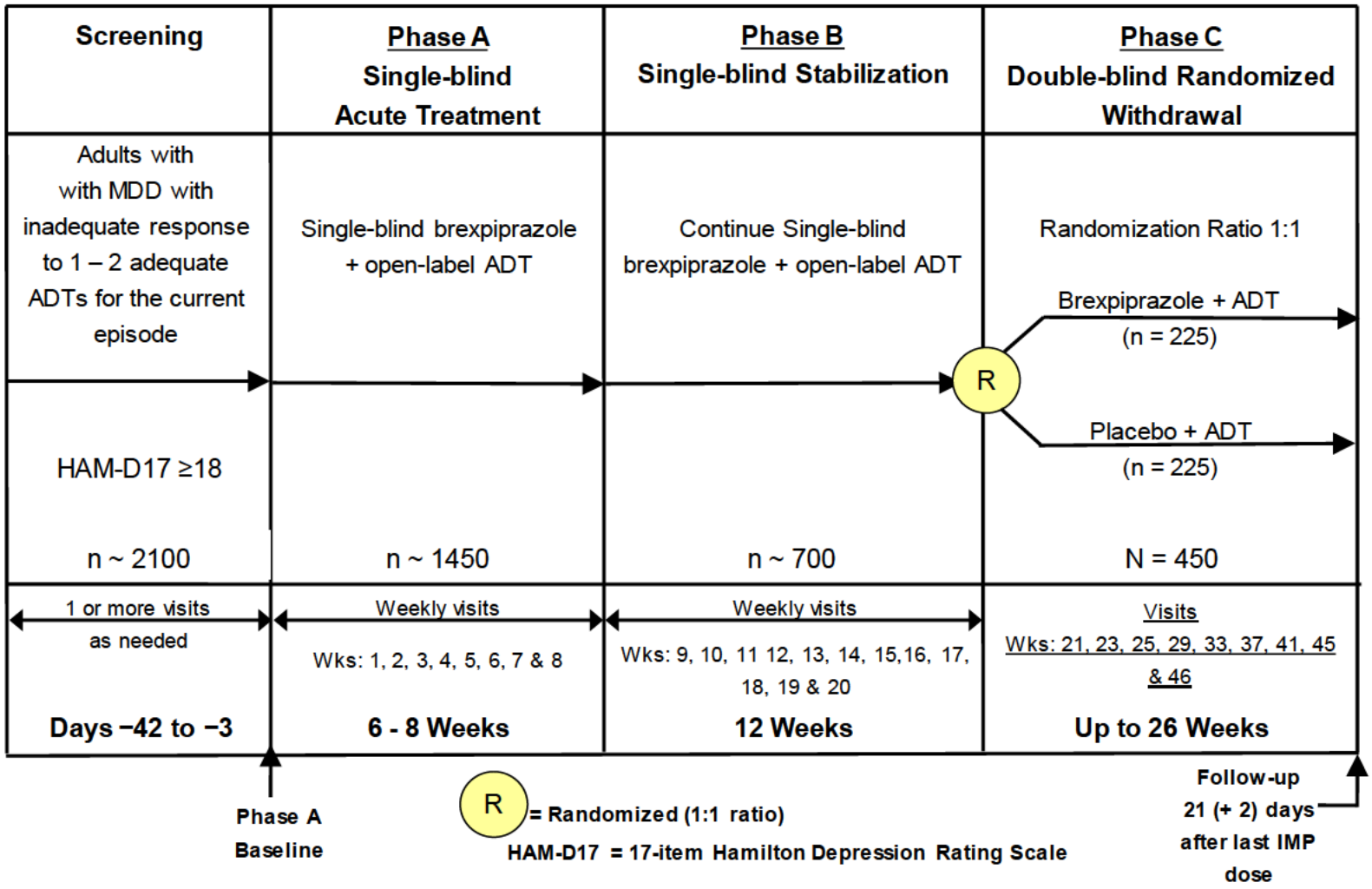


Figure 3-1 Trial Design Schematic

4 Sample Size and Power Justification

The primary objective of this trial is to show superiority of brexpiprazole (2 to 3 mg/day) + ADT over placebo + ADT in time to relapse of MDD signs and/or symptoms. The 2-sided log-rank test will be used to test for statistical significance of the differences between the 2 survival curves.

Based on the results from a completed randomized withdrawal trial of olanzapine/fluoxetine combination versus fluoxetine monotherapy to assess prevention of relapse in subjects with treatment-resistant depression, it is reasonable to expect that 30% of subjects who receive placebo and 16% of subjects treated with brexpiprazole will relapse in 6 months, a hazard ratio of 0.49 (brexpiprazole + ADT versus placebo + ADT) is derived. Thus, 104 relapse events are needed in this trial to reach 95% power to test the primary hypothesis of the protocol at a 2-sided 0.05 alpha level. The sample size estimates used a 1:1 randomization ratio (brexpiprazole + ADT: placebo + ADT) which allowed 2 interim looks at 50% and 75% of events accrual time points. The O'Brian-Fleming boundaries were used for sample size calculation of the interim analysis so that the first interim analysis will be conducted when 52 relapse events are available, and the second interim analysis will be conducted when 78 relapse events are available. The 2-sided alpha levels for these 2 interim analyses are 0.003 and 0.018 respectively, and the alpha left for the final analysis when 104 relapse events are available will be 0.044.

Each subject will be followed for 26 weeks after randomization; allowing for a 20% loss to follow-up, the projected total number of subjects to be randomized into the trial is 450. Using a 1:1 randomization ratio, the number of subjects to be randomized into each treatment group is 225. Assuming approximately 68% of subjects will progress from the Stabilization Phase (Phase B) to Double-blind Randomized Withdrawal (Phase C), it is expected that 700 subjects will be enrolled into the Stabilization Phase of the trial to allow 450 subjects to be randomized. Assuming approximately 55% subjects will progress from the Acute Treatment Phase to the Stabilization Phase, it is expected that approximately 1450 subjects will be enrolled into the Acute Treatment Phase. However, only the number of relapse events (i.e., 104 events) is the determinate factor for sample size of this protocol. The other estimates, such as number of randomized subjects and number of subjects entering the Acute Treatment Phase and the Stabilization Phase, are only based on projection using assumptions that may or may not hold in this trial. Therefore, the actual number of subjects enrolled into the Acute Treatment Phase, the Stabilization Phase, or Double-blind Randomized Withdrawal Phase may vary from the projected values in order to reach the targeted 104 events.

5 Statistical Method

5.1 Data/Data Sets Specifications

5.1.1 Data Sets Analyzed

The following analysis samples are defined for this trial:

- Enrolled Sample: All subjects who sign the informed consents form for the trial and enter Phase A or Phase B.
- Phase A Safety Sample: All subjects who receive at least one dose of brexpiprazole in Phase A.
- Phase B Safety Sample: All subjects who receive at least one dose of brexpiprazole in Phase B.
- Phase B Efficacy Sample: All subjects who enter Phase B (i.e., receive at least one dose of brexpiprazole in Phase B) and have at least one post-baseline efficacy evaluation in Phase B.
- Phase C Safety Sample: All subjects who are randomized to double-blind treatment and receive at least one dose of double-blind trial medication in Phase C.
- Phase C Efficacy Sample: Based on the Intent-to-Treat (ITT) principle, the full analysis set of this trial will be composed of all subjects randomized to the double-blind treatment who take at least one dose of IMP in Phase C.

For the primary analysis of time to event, all subjects belonging to the Phase C Efficacy Sample will be included in the analysis. Subjects who withdraw early from the trial or who are still in the trial at the end of Week 46 will be considered as censoring observations.

The observed case (OC) dataset will consist of the actual observations recorded at each visit and will be used to present summaries per trial week.

The last observation carried forward (LOCF) data set for Phase C will include data recorded at a given Phase C visit or, if no observation is recorded at that visit, data carried forward from the previous Phase C visit. Week 20 data (baseline of Phase C) will not be carried forward to impute missing values for the LOCF data set.

LOCF and OC data sets for Phases A and B will be derived in a manner similar to the process described for Phase C. Baseline of Phase A will not be carried forward to impute missing values for the LOCF in Phase B.

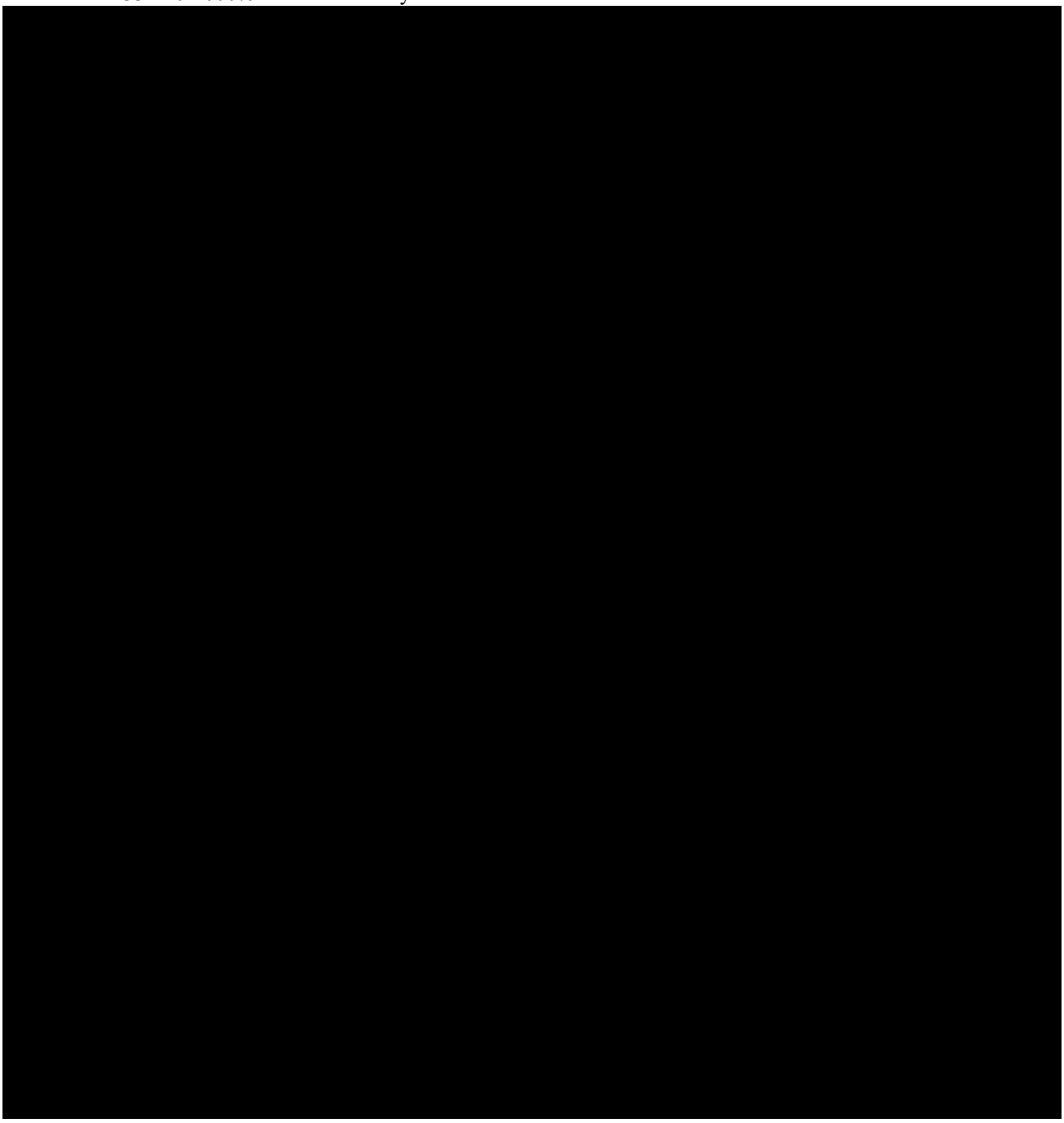
5.1.2 Definition of Baseline and Last Visit for Phase B and C

For analysis purpose, Baseline for Phase A is used as the Phase B baseline. Last visit in Phase B is defined as the last visit with available data prior to randomization, or on/before early termination in Phase B. Baseline for the Double-blind Randomized Withdrawal Phase (Phase C) is Week 20, which is the last visit of the Stabilization Phase prior to the first dose of double-blind IMP in the Double-blind Randomized Withdrawal Phased. If an assessment is missing at Week 20 for a specific measurement, the last none missing value at the Stabilization Phase will be used as the baseline for that assessment; for SDS, the earliest assessment used as phase C baseline cannot be less than Week 14 due to SDS was not accessed weekly. Last visit in Phase C is defined as the last visit with available data at the completion or on/before early termination for the randomized subjects. These are applicable to all efficacy analysis, safety analysis and other outcome analysis.

5.1.3 Handling of Missing Data

For missing data resulting from early withdrawal from the trial in Phase C, subjects who withdraw early will be censored in the primary efficacy analysis at the time when they withdraw from the trial. For other efficacy variables, two types of analyses will be performed for analyses by visit in Phases B and C: LOCF and OC. The LOCF and OC data sets are defined in Section 5.1.1. The OC data set will be used for analyses at each trial visit and the LOCF data set will be used for analyses at the last visit.

The definitions of weeks for Phase A, B and C are described in the following Table 5.1.3-1 and Table 5.1.3-2. Trial Days are derived for each phase from the formula: Trial Day = Date of assessment - Date of 1st brexpiprazole dosing + 1 in Phase A or B, or Date of randomization in Phase C. Based on the number of trial days, subjects are mapped to the corresponding week in each phase in the summary tables. Only last observation within the same week in each phase is used for the summary tables. The LOCF/OC datasets for Phase A consist of data values obtained from visits at Phase A baseline, and week 1 through week 8 in Phase A. The LOCF/OC datasets for Phase B consist of data values obtained from week 1 through week 12 in Phase B and Phase A baseline as the baseline. The LOCF/OC datasets for Phase C consist of data values obtained from visits (after randomization) from week 2 through week 26 in Phase C (study Week 21 – Week 46). Note: NA = Not Applicable.



5.2 Disposition of Subjects

Disposition of subjects for the enrolled sample will be summarized for each phase, while randomized subjects in Phase C (Phase C efficacy sample) will be summarized by treatment group. Completion rate and reason for discontinuation will be tabulated for the enrolled sample in each phase and for randomized subjects by treatment group in Phase C. Subjects who complete Week 46 of Phase C visit are defined as the completers. Subjects who meet the stability criteria (as specified in Section 3.3) for the consecutive

12 weeks in Phase B will be summarized for the enrolled sample and randomized sample by week in Phase B.

5.3 Demographic and Baseline Characteristics

Baseline and demographic characteristics will be summarized using descriptive statistics (frequency, mean, median, standard deviation (SD), maximum, minimum, and percentage when applicable). These characteristics include age, race, ethnicity, gender, weight, height, body mass index (BMI), and waist circumference will be tabulated for the enrolled sample by phase and for the randomized subjects by treatment group. Baseline disease severity including baseline , MADRS, Clinical Global Impression - Severity of Illness scale (CGI-S) score, SDS, will be summarized for the enrolled sample in each phase and for the randomized subjects by treatment group. Baseline total 17-item Hamilton Depression Rating Scale (HAM-D17) will be summarized for enrolled sample. In addition, age at first diagnosis of MDD will be summarized. If both month and date are missing for the first diagnosis date, June 30 is used in calculating age at first diagnosis of MDD. If only date is missing for first diagnosis date, date 15 is used. If only month is missing for first diagnosis date, June is used as the month. If year of first diagnosis date is missing, age at first diagnosis of MDD is missing.

Disease severity and psychiatric history at Phase C baseline (or at screening) will also be summarized by descriptive statistics for the Safety Sample to identify any potential lack of balance between the treatment groups for the Phase C efficacy sample.

5.4 Protocol Deviations

Protocol deviations will be summarized by center and type of deviation for randomized subjects by treatment group. A listing of protocol deviations will be provided.

5.5 Efficacy Analysis

5.5.1 Primary Efficacy Analysis

5.5.1.1 Primary Efficacy Endpoint

The primary efficacy endpoint of this trial is time-to-relapse by any criteria as defined in the blinded addendum to this protocol, measured from Week 20 (randomization) to the Double-blind, Randomized Withdrawal Treatment Phase).

Time-to-relapse (primary efficacy endpoint) is defined as meeting any of the following criteria (blinded to investigators) measured from randomization into Phase C:

I. At the same visit, an increase on the MADRS total score of 50% from randomization and CGI-S score ≥ 4, **OR**

II. Hospitalization for depression, **OR**
III. Discontinuation for lack of efficacy or worsening of depression, **OR**
IV. Active suicidality, defined as a score of ≥ 4 on the MADRS item 10 **OR** an answer of "yes" on question 4 or 5 on the C-SSRS **OR** an answer of yes to any of the questions on the Suicidal Behavior section of the C-SSRS.

The time origin for measuring this event time is the date of randomization into the double-blind, placebo-controlled treatment phase (Phase C). Phase C Efficacy Sample is used in the primary analysis of the primary efficacy endpoint. The primary objective of statistical analysis is to compare the efficacy of brexpiprazole (2 to 3 mg/day) + ADT with that of placebo + ADT with regard to time-to-relapse. This will be achieved by testing the equality of the two survival curves by the log-rank test.

Interim analyses are planned to be performed at approximately 50% and 75% of events accrual time points using the O'Brian-Fleming boundaries for rejection of the null hypothesis. Details of the interim analyses are provided in the interim analysis plan.

A group sequential testing procedure will be employed with two planned interim analyses at approximately 50% (52 events) and 75% (78 events) events accrual time points, and a final analysis using 100% (approx. 104 events) events. The O'Brian-Fleming boundaries will be utilized in this analysis for the rejection of the null hypothesis to maintain an overall nominal significance level of 0.05 (two-sided). [redacted]

[redacted]

While the interim analysis follows the boundaries listed in Table 5.5.1.1-1, "final" analysis must be performed after a recommendation based on the interim analysis. This "final" analysis will include the relapse events that occur between the data cutoff for the interim analysis and the final database lock, and the alpha level for the "final" analysis will be derived based on the total number of events using O'Brien-Fleming spending function. Should the interim analysis be performed at a different time, different frequency, the spending function approach as originally introduced by Lan and DeMets will be applied to adjust the critical values to guarantee an overall 0.05 alpha level. The

spending functions corresponding to O'Brien-Fleming boundaries will be formally applied to the time-to-event analysis (using the log-rank test).

For the primary endpoint, a 95% confidence interval for the hazard ratio (brexpiprazole (2 to 3 mg/day) + ADT with that of placebo + ADT) will be provided using the Cox Proportional Hazard model with treatment as the fixed effect in the model.

The Phase C Efficacy Sample will be used in the above analysis. Subjects who withdraw early from the trial without relapse events or who are still in the trial at the end of the Week 46 will be considered as censored observations at the discontinuation or completion date. If the time to event of the subject or completion date is greater than 196 days from randomization, then the subject will be censored at 196 days from randomization with no event. Subjects who stay in the trial but have no relapses in an interim analysis will be censored at the date of data cut-off for the interim analysis.

5.5.1.2 Technical Computational Details for Primary Efficacy Analysis

Subjects who meet relapse criteria are considered as an event of relapse. For relapsed subjects in Phase C, time to event is calculated as the earliest date of time-to-relapse, defined as the time meeting relapse criteria - randomized date + 1. Below are the criteria to determine the relapse time for each relapse criteria:

1). If subjects meet the relapse criteria per MADRS and CGI-S assessments of criteria I, which are conducted on different dates, the later assessment date will be used as the relapse event date.

2). If subjects meet the relapse criteria per criteria IV, C-SSRS or MADRS, the earliest date of the evaluation that the criteria are met will be used as the relapse event date.

3). If subjects meet the relapse criteria per hospitalization due to worsening of depression and are reported as an adverse event (AE), the AE start date will be used as the relapse event date.

4). If subjects meet the relapse criteria per lack of efficacy or worsening of depression but are not reported as an AE, the first two criteria stated prior will be checked. The earliest date that criteria 1) or 2) is met will be the relapse event date. If none of the above is applicable, then the discontinuation date will be the relapse data.

Subjects who complete the double-blind treatment period of 26 weeks or discontinue treatment without meeting the relapse criteria (including those who are lost to follow-up during treatment period) in Phase C will be considered as censored observations as

mentioned above. For the withdrawn (rather than lack of efficacy) subjects in Phase C, they are considered as censoring at discontinuation, i.e., time to censor = discontinuation date - randomization date + 1. For the completed subjects, they are considered as censoring at the completion, i.e., time to censor = completion date - randomized date + 1. If the time to event/censor is great than 196 days, the subject, no matter discontinued or completed Phase C, will be considered as censoring at Day 196.

The log-rank test will be used to test the equality of the two survival curves. Additionally, a nominal 95% confidence interval for the hazard ratio (brexpiprazole vs. placebo) will be provided using the Cox Proportional Hazard model with terms for treatment in the model. Ties in the event times will be handled by using the exact option in SAS PROC PHREG.

5.5.2 Sensitivity Analysis for the Primary Endpoint

In order to assess the effect of potential informative censoring prior to the scheduled 26 weeks of treatment in Phase C, two approaches of sensitivity analysis will be adopted to estimates this effect on the primary efficacy analysis results. These are: (1) multiple imputation of discontinued censored observations using the estimated hazard rate from the placebo group, (2) treating 20% of randomly selected discontinued patients only from the brexpiprazole group as events. These methods of sensitivity analyses are described below. Subjects who discontinued in Phase C because of trial termination by the sponsor will not be imputed in the final analysis.

5.5.2.1 Multiple Imputation Method

Assume that the time to event variable follows an exponential distribution. The following multiple imputation method will be used to impute time to event from their time of discontinuation for censored subjects in the sensitivity analysis.

For a placebo + ADT subject, assume the hazard of the subject to develop an event is identical to the estimated hazard of the placebo + ADT group (hazard estimated at interim analysis for sensitivity analysis of the interim analysis and hazard estimated in the final analysis for sensitivity analysis of the final analysis). Event time will be simulated for the subject using an exponential distribution with the rate parameter (λ) equal to the estimated hazard of the placebo group. This same rate is used to simulate Event time of all dropout subjects from both brexpiprazole and placebo treatment groups; except that the placebo group will use the rate without modification and the brexpiprazole group will use the rate multiply by a factor equal to or greater than 1. The time to event of the subject is equal to the sum of the time to discontinuation and the simulated event time of the subject. If the time to event of the subject is greater than 196 days from

randomization, then the subject will be censoring at 196 days from randomization with no event. Otherwise, the subject will have an event at the time to event as calculated above.

For a brexpiprazole + ADT treated subject, the imputation of time to event or time to censor (if the simulated event is greater than 196 days) is similar to the imputation of placebo treated subject, except that the exponential distribution will have a rate parameter equal to k* λ, where λ is the hazard rate of the placebo group and k $\geq$1.

For each imputation, the simulated data will be combined with the data for subjects who complete the double-blind treatment period of 6 months without an event and subjects with events, to obtain the log-rank test and in the time to event analysis using Cox Proportional Hazard model with treatment as the independent variable. For each k, 30 imputations will be performed. The results of these thirty imputations will be combined using SAS PROC MIANALYZE to provide a final result for this k. The k will run from 1.0 to 1.1, 1.2, …, until the test loses its significance.

5.5.2.2 Discontinuations as Events for Brexpiprazole Group

The second sensitivity analysis will be conducted by considering 20% randomly selected discontinued subjects (other than subjects discontinued because of study termination by the sponsor) in Phase C on Brexpiprazole treatment without meeting the relapse criteria as having the relapse at one day after the discontinuation date. Then the Cox Proportional Hazard model with treatment as the fixed effect will be generated.

For relapsed subjects in Phase C, time to event will be calculated as the date of meeting criteria - randomized date + 1. For the 20% selected discontinued subjects in the brexpiprazole treatment group without meeting the relapse criteria in Phase C, time to event will be calculated as the date of discontinuation - randomized date + 2. For the other 80% discontinued subjects on brexpiprazole treatment group without meeting relapse criteria in Phase C and all discontinued subjects on placebo treatment group without meeting relapse criteria in Phase C, they will be considered as censored observations at the discontinuation date. Time to censoring will be calculated as the date of discontinuation - randomized date + 1. Subjects who complete the double-blind treatment period of 6 months in Phase C will be considered as censoring observations. Time to censoring will be calculated as completion date - randomization date + 1. In total, 30 imputations will be made for this sensitivity analysis. Again, SAS PROC MIANALYZE will be used to combine the results of these 30 imputations to provide a final result.

5.5.3 Effect of COVID-19 on the Primary Endpoint

In earlier 2020, an outbreak of respiratory disease caused by a novel coronavirus named "Coronavirus Disease 2019" (COVID-19) had widely spread all over the world. On March 13, 2020, the President of the United States declared a national emergency in response to COVID-19. It is after this date, centers enrolled in this study had implemented remote visits in response to this pandemic. In order to assess the effects of this change on the primary efficacy endpoint, subgroup analysis is planned to assess the effect of COVID-19 on the primary analysis.

The COVID subgroup is defined for subjects enrolled in Phase C and classified according to the following criteria: all subjects randomized into Phase C after March 13th will start the remote access or remote changes for their visits and will be classified as COVID=Yes; subjects randomized on or before March 13th 2020 will have COVID=No. The COVID-19 related changes will include but not limit to remote visit, missed visit (especially for efficacy assessments), missed assessment, discontinuation due to COVID-19.

A Cox Proportional Hazard model with treatment as the fixed effect will be fitted by each of the two COVID subgroups. Subjects who are randomized on or before March 13th , 2020 will be censoring on March 13th 2020 or the last in-person visit whichever is first. This way the analysis for the COVID=No subgroup will estimate the hazard ratio of the two treatment groups without any impact from the COVID. The within subgroup p-value from the log-rank test from the treatment comparison will be provided for each subgroup, together with the hazard ratios and their 95% confidence intervals using the Cox Proportional Hazard model with treatment as a fixed effect in the model. A listing of all participants affected by the COVID-19 related study disruption/changes will be presented with a description of how the individual's participation was altered.

5.5.4 Secondary Efficacy Analysis

5.5.4.1 Secondary Efficacy Analysis

The secondary efficacy endpoints evaluated for Phase C will include:

- Time-to-Functional relapse
- Proportion of subjects meeting relapse criteria
- Proportion of subjects maintaining remission
- Change from baseline to endpoint in MADRS total score (LOCF)
- Change from baseline to endpoint in CGI-S score (LOCF)
- CGI-I score (LOCF)

- Change from baseline in SDS mean total score and each of the SDS individual item scores (LOCF)

5.5.4.2 Analysis of Functional Relapse

The secondary efficacy endpoint is time-to-functional relapse based on a 30% increase in Sheehan Disability Scale (SDS) mean total score from Phase C baseline and at least one SDS sub-score at 4 or greater and a SDS total score great than or equal to 7 when all 3 sub-scores are available (or SDS total score greater than or equal to 5 when work/school item does not apply) at Phase C visits. In the event that the SDS phase C baseline is 0, the evaluation will be based on the total SDS and SDS sub-score criteria at Phase C. Time to functional relapse will be the first SDS assessment in Phase C that met the functional relapse criteria.

The Sheehan Disability Scale (SDS) is a self-rated instrument used to measure the effect of the patient's symptoms on work/school, social life, and family/home responsibilities. For each of the three items, scores range from 0 through 10. The number most representative of how much each area was disrupted by symptoms is marked along the line from 0 = not at all, to 10 = extremely. For the work/school item, no response will be entered if the patient did not work or go to school for reasons unrelated to the disorder and a response therefore not being applicable. The Mean total SDS Score will be calculated over the three item scores. All three item scores need to be available with the exception of the work/school item score when this item is not applicable. In the event that work/school item does not apply, the other two items will be averaged as the mean total score.

Event/censer time will be defined as the following: if subjects meet the relapse criteria for functional relapse, the SDS assessment date during Phase C will be used as an event date; If the subject withdraw or discontinue study without meeting the functional relapse criteria, the withdraw/discontinuation date will be the censoring date. For completed subjects, they are considered as censoring at the completion, i.e., time to censor = completion date - randomized date + 1. If time to event/censor is great than 196 days, this subject no matter if he/she discontinued or completed for Phase C will be considered as censoring at Day 196. At interim analysis, the interim analysis cut-off date will be used as the censor date. Log-rank test will be used for the time-to-functional relapse comparisons similar to the primary efficacy analysis.

5.5.4.3 Analysis of Other Secondary Efficacy Endpoints for Phase C

Baseline for Phase C is defined as the last visit of Phase B prior to the first dose of double-blind trial medication in Phase C. Analysis of the change from baseline for other secondary efficacy endpoints will be based on subjects in the Phase C Efficacy Sample. Analysis of the CGI-S score endpoint will be based on subjects in the Phase C Efficacy Sample who have post-baseline CGI-S observations. The analysis will be based on all subjects who have been randomized and taken one dose of IMP in Phase C.

The proportion of subjects meeting relapse criteria as defined in Primary Efficacy Endpoint Section 5.5.1.1 will be analyzed by treatment group, visit (OC) and the subject's last visit (LOCF) in Phase C using the Chi-square test.

Subjects maintaining remission is defined as MADRS total score ≤ 10. Analysis will be done at each visit (OC) and the final visit (LOCF) using Chis-square test.

The MADRS, CGI-S score and individual SDS score at endpoint will be analyzed at the last visit (LOCF) using an ANCOVA model. As an exploratory analysis, change from baseline variables including MADRS, CGI-S, and SDS will be analyzed by a Mixed Model Repeated Measures (MMRM) approach using observed data, with terms of treatment, trial center, visit, treatment-by-visit interaction, baseline and baseline-by-visit interaction. To avoid too many parameters in the statistical model, variance-covariance matrix with heterogeneous compound symmetry structure is assumed in the MMRM analysis. In addition, the count and frequency of SDS with work/school item not applicable at baseline will be summarized.

The CGI-I score at endpoint will be analyzed at the last visit of LOCF data using the Cochran-Mantel-Haenszel method based on raw mean score statistics.

Kaplan-Meier curves will be plotted for time to discontinuation due to all causes and the log-rank test will be used to test the differences in the survival curves.

5.5.4.3.1 Analysis of Other Secondary Efficacy Endpoints for Phase B and Phase A

Descriptive statistics will be provided by visit for change from baseline to End of Phase B in MADRS Total Score, CGI-S, SDS scores; analysis will be provided by visit in Phase B. These analyses will be based on OC data at each Phase B visit and the LOCF data at the last visit of Phase B, using Phase B Efficacy Sample. Baseline for Phase B will be the same as the baseline for Phase A for all above efficacy endpoint. Time to discontinuation due to all causes will also be examined in Phase B starting from the first dose of study medication taken in Phase B to discontinuation or completion.

Efficacy scales assessed during Phase A will be provided in listings for the Phase A Safety Sample.

5.5.4.3.2 Technical Computational Details for Other Secondary Efficacy Analysis

1) In general, data for MADRS, CGI-S, and SDS assessment after the relapsed dates for the relapsed randomized subjects will not be included in the Phase C efficacy summary tables.
2) Time to discontinuations due to all causes rather than the sponsor discontinued trial will be measured from the date of randomization to date of early termination for withdrawn subjects in Phase C, i.e., time to discontinuation = discontinuation date - randomization date +1. Subjects discontinuing the trial due to the sponsor discontinued trial in Phase C will yield censoring observations for time to discontinuation and be considered censoring at discontinuation, i.e., time to censor= discontinuation date - randomization date +1. Subjects completing the entire treatment duration of Phase C will yield censoring observations for time to discontinuation and be considered censoring at completion, i.e., time to censor= completion date - randomization date +1. If time to discontinuation/completion is great than 196 days, the subject will be considered as censoring at Day 196, no matter if he/she discontinued or completed Phase C.
   Time to discontinuation due to all causes rather than the sponsor discontinued trial in Phase B will be calculated for Phase B efficacy sample using the similar method. Time to discontinuations due to all causes rather than the sponsor discontinued trial in Phase B will be measured from the 1st dosing date of Phase B to date of early termination for withdrawn subjects in Phase B, i.e., time to discontinuation in Phase B = discontinuation date - 1st IMP dosing date of Phase B +1. Subjects discontinuing the trial due to the sponsor discontinued trial in Phase B will yield censoring observations for time to discontinuation and be considered censoring at discontinuation, i.e., time to censor= discontinuation date - 1st IMP dosing date of Phase B +1. Subjects continuing into Phase C will yield censoring observations at randomization date, i.e., time to censor in Phase B = randomization date - 1st IMP dosing date of Phase B +1. If time to discontinuation/censor is great than 98 days, this subject will be considered as censoring at day 98 no matter if she/he is randomized into Phase C.
3) Computing of Scale Scores
   a) The CGI-S and CGI-I of 0 will be set to missing because value means ‘not assessed’.

The MADRS consists of 10 items, all rated on a scale of 0 to 6 with 0 being the “best” rating and 6 being the “worst” rating. The MADRS Total Score is the sum of ratings for all 10 items. The MADRS Total Score will be un-evaluable if more than 2 items are missing. If 8 or 9 of the 10 items are recorded, the MADRS Total Score will be the mean of the recorded items multiplied by 10 rounded to the first decimal place.

The CGI consists of two scales: CGI Severity (CGI-S), and CGI Improvement (CGI-I). CGI-S items are: 0 = not assessed; 1 = normal, not at all ill; 2 = borderline mentally ill; 3 = mildly ill; 4 = moderately ill; 5 = markedly ill; 6 = severely ill; 7 = among the most extremely ill patients. The score 0 (= not assessed) will be set to missing. The CGI-S is therefore a 7-point scale from 1 through 7.

5.5.5 Sub-group Efficacy Analysis

Sub-group analysis of the primary endpoint in Phase C will be provided by region (US and non-US), gender (male and female), race (Caucasian and non-Caucasian), and age (<45 years old and ≥45 years old at Phase C baseline). Subgroup analyses will be performed for the FAS. To explore the beneficial effects of brexpiprazole vs placebo in the subgroups, the estimated hazard ratio, two-sided nominal 95% confidence interval, and within subgroup p-value from the log-rank test will be provided for each of the subgroups based on the Cox's proportional hazards model in which treatment is included as the fixed-effect factor. It should be aware that positive findings from these subgroup analyses have to be interpreted with caution since there is a non-negligible chance of false positives.

5.5.6 Exploratory Efficacy Analysis

The number and percentage of subjects meeting the below the specified criteria are summarized by visit:

- Subjects responding to treatment during Phase A by visit as specified in Section 3.2.
- Subjects meeting stabilization criteria of MADRS and CGI-S at any point during Phase B Section 3.3.
- Subjects achieving remission at any time point during Phase B.

Subjects achieving remission in Phase B is defined as a MADRS total score ≤ 10.

5.6 Safety Analysis

Standard safety variables to be analyzed include AEs, clinical laboratory tests, vital signs, ECGs, body weight, waist circumference, and BMI. In addition, data from the following safety scales will be evaluated: Simpson-Angus Scale (SAS) Total Score, Abnormal Involuntary Movement Scale (AIMS) Total Score, Barnes Akathisia Rating Scale (BARS), and C-SSRS.

In general, safety analysis will be summarized for Phase B safety sample and Phase C safety sample respectively. Listing of safety variables for Phase A safety sample will be

provided. Baseline for Phase A safety sample will be used as the baseline for safety analysis in Phase B as well. Baseline for Phase C safety sample is defined as the last visit with available data in Phase B prior to the first dose in Phase C.

5.6.1 Adverse Events

All adverse events (AE) will be coded by MedDRA System Organ Class (SOC) and Preferred Term (PT). A treatment-emergent AE (TEAE) is defined as an AE which starts after start of trial medication), or an AE continues from baseline of the specific phase and becomes serious, worsening, trial drug-related or results in death, discontinuation, interruption or reduction of trial medication during this phase. The incidences of the following treatment-emergent adverse events (TEAEs) will be summarized by treatment groups:

- TEAEs by severity
- Potentially drug related TEAEs
- TEAEs with an outcome of death
- Serious TEAEs
- Discontinuations due to TEAEs

Listing and incidences of AEs above will be tabulated by treatment for Phase B safety sample and Phase C safety sample, respectively. Deaths, SAEs, and AEs leading to discontinuation from trial will be listed by phase for the enrolled sample.

5.6.2 Laboratory Test Results

Clinical laboratory tests include the routine clinical laboratory measurements, prolactin concentrations, coagulation parameters (PT, aPTT, and INR), HbA1c, cortisol, ACTH, and TSH as well as laboratory tests identified using prospectively defined criteria. Summary statistics for the clinical laboratory measurements at baseline and post-baseline visits, and summary statistics of changes from baseline to each visit will be presented by treatment group. Incidence of treatment-emergent potentially clinically significant abnormal lab results will also be summarized by treatment groups. Listings of potentially clinically significant abnormalities will also be provided. Criteria of potentially clinically significant lab test abnormalities are provided in Appendix 1.

According to FDA Guidance, laboratory measurements that signal the potential for drug-induced liver injury (DILI) will be reported. An incidence table and a listing will be provided for subjects who meet one or combinations of following criteria:

1) ALT or AST ≥ 3× upper limit of normal (ULN) or baseline value

2) increase in bilirubin ≥ 2× ULN or baseline value

If laboratory tests assessments are repeated for the same visit, the last repeated values are used for summary tables. This is accomplished by sorting patient data by visit date and visit time (if applicable) within the same visit identification. If the lab data are recorded as ranges (i.e., including < or > limit of quantification), these data are not included in the calculations for changes from baseline but included in the calculations for incidences.

Listings and summary tables of clinical laboratory measurements will be presented for Phase B safety sample and Phase C safety sample respectively. Potentially clinically significant laboratory abnormalities for Phase A safety sample will be provided by listings.

5.6.3 Vital Signs Data

Vital signs include body temperature, heart rate, systolic blood pressure, and diastolic blood pressure. In addition, body weight, waist circumference and body mass index (BMI) are measured. Descriptive statistics will be provided by treatment group, for change from baseline in vital signs parameters. Incidence of treatment-emergent potentially clinically significant vital sign results will also be summarized by treatment groups. Listings of potentially clinically significant abnormalities will also be provided. Criteria for the potentially clinically significant vital sign abnormalities are provided in Appendix 2.

If vital sign assessments are repeated for the same visit, the last repeat values will be used for production of mean change from baseline. This is accomplished by sorting patient data by visit date and visit time (if applicable) within the same visit identification.

Listings and summary tables of vital sign data will be presented for Phase B safety sample and Phase C safety sample. Potentially clinically significant abnormalities in vital sign will be listed by subjects for Phase A safety sample.

5.6.4 ECG Data

For the analysis of QT and QTc, data from three consecutive complexes (representing three consecutive heart beats) will be measured to determine average values. The following QT corrections will be used for reporting purposes in the clinical trial report:

1) QTcB is the length of the QT interval corrected for heart rate by the Bazett formula: $QTcB=QT/(RR)^{0.5}$
2) QTcF is the length of the QT interval corrected for heart rate by the Fridericia formula: $QTcF=QT/(RR)^{0.33}$

3) QTcN is the length of the QT interval corrected for heart rate by the FDA Neuropharm Division formula: $QTcN=QT/(RR)^{0.37}$

The potentially clinically significant ECG abnormalities will be listed by subject. The incidences of abnormal ECGs of potential clinical significance will be tabulated. Descriptive statistics of change from baseline in heart rate and ECG intervals will be provided. Criteria for potentially clinically significant ECG abnormalities are provided in Appendix 3. The listings and summary tables of ECG abnormalities will be provided for Phase B safety sample and Phase C safety sample, and only the listings will be provided for Phase A safety sample.

In summarizing the incidence of abnormalities, a patient must have had an evaluation that met abnormality criteria by the end of trial, i.e., last contact date. Incidence rate is calculated as the number of patients having at least one abnormality within a trial period divided by the number of patients who are both exposed to the trial medication and have an on-treatment evaluation within that trial period.

5.6.5 Physical Examination

Physical examination findings will be listed by subject for each trial phase.

5.6.6 Extent of Exposure

5.6.6.1 Extent of Exposure to Brexpiprazole and Placebo

The start date of double-blind study therapy - brexpiprazole or placebo - will be the first day of double-blind dosing. The number and percentage of patients who receive brexpiprazole/placebo will be presented by phase, week, and treatment group for the Safety Sample of Phase C. This summary will be performed for Enrolled Sample for Phases A and B.

The mean daily dosage will be summarized by visit and treatment group using descriptive statistics. The mean daily dosage per patient per week will be determined for each week of the study. This will be calculated by dividing the sum of individual total doses by the number of days in the week interval. The summary will contain for each treatment group the number of patients receiving double-blind study medication, and the mean and range of the mean daily dose for each week.

5.6.6.2 Extent of Exposure to ADT

The number and percent of patients who receive an ADT in Phase C will be presented by visit, by ADT treatment group, and by double-blind treatment group for the Safety Sample. A similar summary will be prepared for Phase B. For each ADT, the mean daily dose will be calculated in the same way as for double-blind study medication.

Summaries will be presented for ADT administered to patients in the ADT Sample during Phase A and Phase A+ Sample during Phase A+.

In addition, the mean daily dosage of ADT will be summarized for the ADT Sample during Phase A and the Phase A+ Sample during Phase A+ by week and by ADT group using descriptive statistics.

5.6.7 Other Safety Data Analysis

5.6.7.1 Suicidality Data

Suicidality monitored during the trial using the C-SSRS will be summarized as the number and percentage of subjects reporting any suicidal behavior, ideation, behavior by type (4 types), ideation by type (5 types), and treatment-emergent suicidal behavior and ideation.

Summary tables of C-SSRS will be provided for Phase B safety sample and Phase C safety sample. Suicidality is defined as report of at least one occurrence of any type of suicidal ideation or at least one occurrence of any type of suicidal behavior during assessment period (count each person only once).

Treatment emergent suicidal behavior and ideation is summarized by four types: Emergence of suicidal ideation, Emergence of serious suicidal ideation, Worsening of suicidal ideation, Emergence of suicidal behavior.

Emergence of suicidal behavior/ideation is defined as report of any type of suicidal behavior/ideation during treatment when there was no baseline suicidal behavior/ideation.

Emergence of serious suicidal ideation is defined as observation of suicidal ideation severity rating of 4 or 5 during treatment when there was no baseline suicidal ideation.

Worsening of suicidal ideation is defined as a suicidal ideation severity rating that is more severe than it was at baseline.

5.6.7.2 EPS Rating scales

5.6.7.2.1 Analysis of EPS Rating Scales

Descriptive statistics will be provided for change from baseline to end of phase in SAS, AIMS, and BARS scores for Phases B and C safety samples. Results will be summarized by visit. In addition, change from end of Phase B to end of Phase C in scores for the SAS, AIMS (total of the first 7 item scores), and BARS scales will be evaluated using ANCOVA with baseline value as covariate and treatment and center as factors. OC data sets will be used in the analyses of these EPS scales. In addition, analyses will be

performed using the maximum (i.e. the worst) value observed during Phase C and the last visit data to determine the change from baseline score. The ANCOVA model for change at the last visit and for change to the maximum value will include the Phase C baseline, study center and treatment group. The same analyses will be performed on the AIMS individual item scores 8, 9, and 10. In addition, incidence of BARS Global Clinical Assessment of Akathisia during Phase B and C by severity category will be provided. Analyses of these EPS rating scales will be performed for the Safety Sample.

5.6.7.2.2 Technical Computational Details for EPS Rating Scales

1) The SAS total score is the sum of the rating scores for 10 items from the SAS panel in the CRF. The SAS Total Score is un-evaluable if less than 8 of the 10 items are recorded.
   If 8 or 9 of the 10 items are recorded, the Total SAS score is the mean of the recorded items multiplied by 10 and then rounded to the first decimal place.
2) The AIMS movement rating score is the sum of the rating scores for facial and oral moments (i.e., item 1 - 4), extremity movements (i.e., item 5 - 6), and trunk movements (i.e., item 7). The AIMS Total score is un-evaluable if less than 6 of the 7 items are recorded. If 6 of the 7 items are recorded, the Total Score is the mean of the recorded items multiplied by 7 and then rounded to the first decimal place.
3) The BARS score is based only on the item of Global Clinical Assessment of Akathisia.

5.7 Other Outcome Analysis

[redacted]

5.8 Pooling of small centers

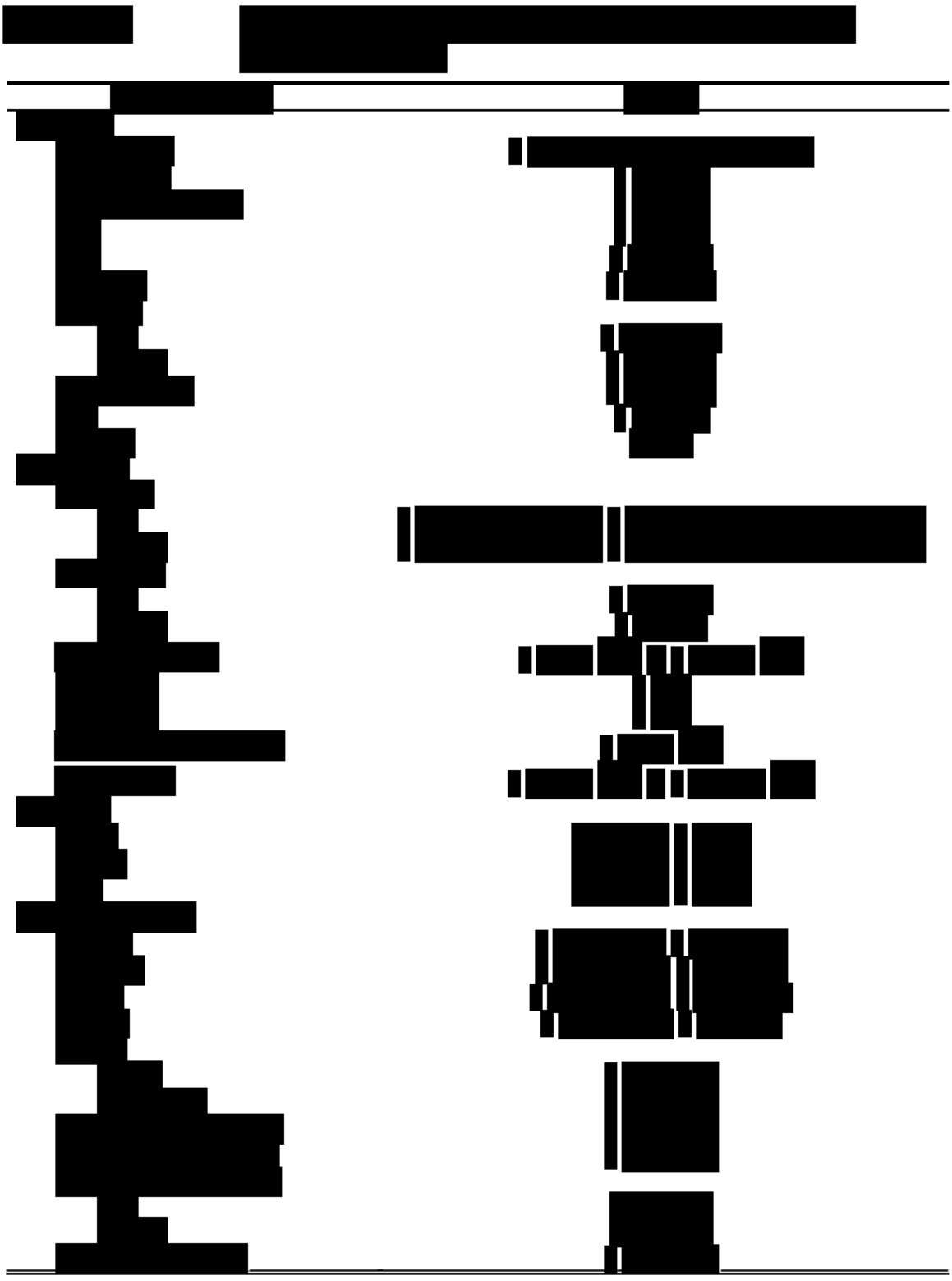
In models such as ANCOVA or MMRM with center included as fixed effect, small centers will be pooled. Small centers will be defined as centers that do not have at least one evaluable subject (evaluable with regard to the primary efficacy variable) in each treatment arm in Phase C. All small centers will be pooled to form “pseudo centers” for the purpose of analysis according to the following algorithm. Small centers will be ordered from the largest to the smallest based on the number of evaluable subjects (i.e.,

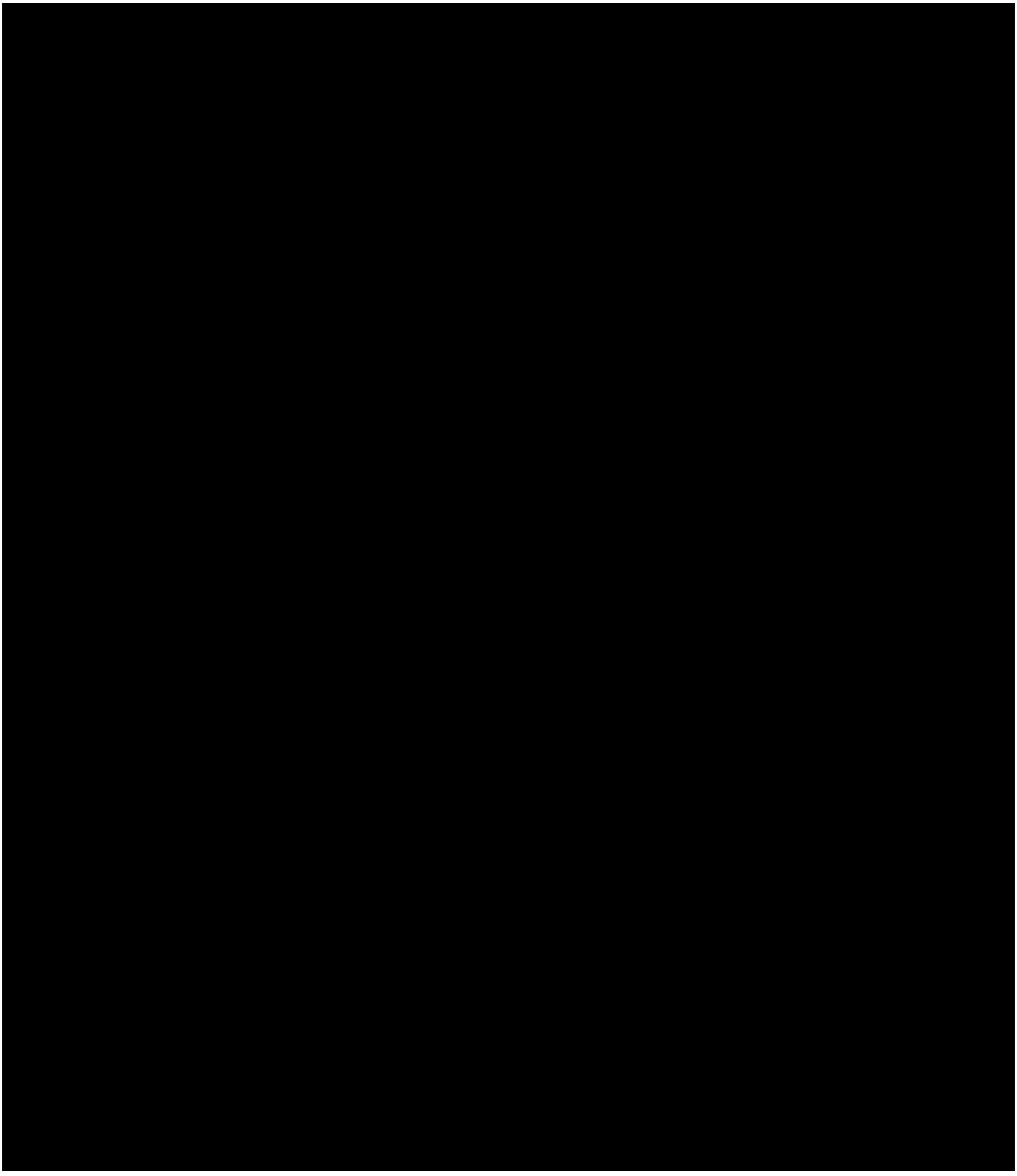
subjects who have a baseline value and at least one post-randomization value. The process will start by pooling the largest of the small centers with the smallest of the small centers until a non-small center is formed. This process will be repeated using the centers left out of the previous pass. In case of ties in center size, the center with the smallest center code will be selected. If any centers are left out at the end of this process, they will be pooled with the smallest pseudo centers, or if no pseudo centers exist, they will be pooled with the smallest non-small center.

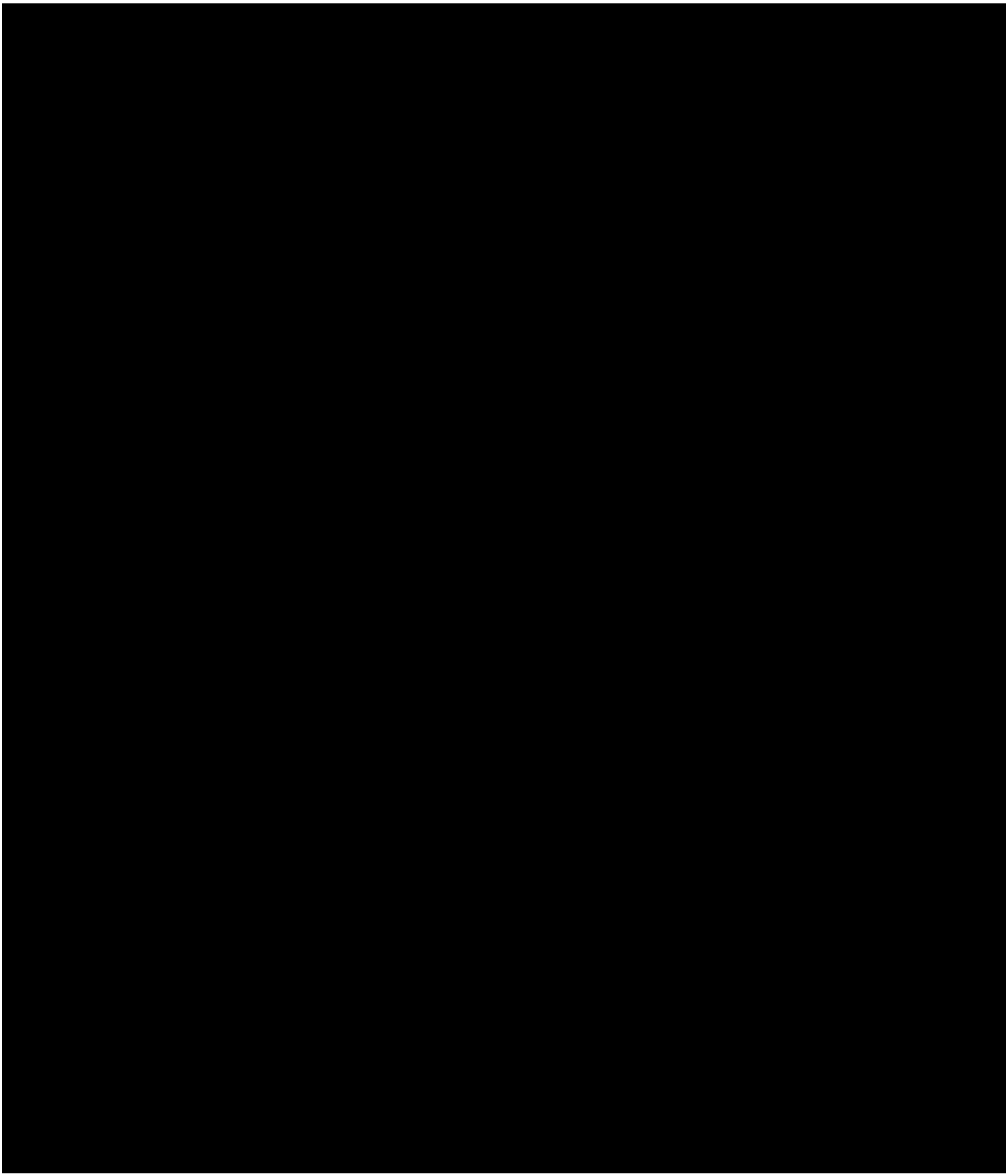
[redacted]

6 References

E. Glimm, W. Maurer and F. Bretz Hierarchical testing of multiple endpoints in group-sequential trials. *Statist. Med.* ***2010****,* 29 219—228.







[redacted]