NCT4414930: Pharmacologic augmentation of targeted cognitive training in schizophrenia
Statistical Analysis Plan

October 11, 2019

Statistical Design & Power

Design (see "Research Strategy", Figure 7): Screened, eligible patients complete clinical, neurocognitive and functional measures and candidate biomarkers. In Test 1, all subjects are tested in Sound Sweeps after PBO, and then assigned to TCT+PBO vs. TCT+AMPH arms (n=27/arm) using stratified random sampling (over sex, age and high/low Test 1 (baseline) APS learning) blind to arm identity, similar to[107]. Stratifying for baseline APS learning should increase the sensitivity of the primary target engagement metric. An example of a table for arm assignment is seen here:

Gender	Baseline	Age	Group	Group	Group	Group	Group	Group	Group	Group
Male	High	< 39	PBO	PBO	AMPH	AMPH	AMPH	AMPH	PBO	PBO
	High	≥ 39	PBO	AMPH	PBO	AMPH	AMPH	PBO	AMPH	PBO
	Low	< 39	AMPH	AMPH	PBO	PBO	PBO	PBO	AMPH	AMPH
	Low	≥ 39	AMPH	PBO	AMPH	PBO	PBO	AMPH	PBO	AMPH
Female	High	< 39	PBO	PBO	AMPH	AMPH	AMPH	AMPH	PBO	PBO
	High	≥ 39	PBO	AMPH	PBO	AMPH	AMPH	PBO	AMPH	PBO
	Low	< 39	AMPH	PBO	AMPH	PBO	PBO	AMPH	PBO	AMPH
	Low	≥ 39	AMPH	AMPH	PBO	PBO	PBO	PBO	AMPH	AMPH

Example of table for stratified random sampling, by gender, baseline APS learning (post-PBO) and age (based on approximate mean age (39 yo) of last 100 SZ patients tested in our program)

Test 2 follows approximately 5-7 days later. Tests 1 and 2 are used to assess target engagement (Aim 1: AMPH-enhanced APS learning) and AMPH effects on auditory discrimination. Our design allows for 20% attrition from enrollment (n=69) to completion of target engagement testing (n=54).

For Aim 2, TCT is scheduled 3 d/week and continues until a subject completes 30-h (approximately 10-12 weeks). 60 min prior to each TCT session, patients take either PBO or 5 mg AMPH, as per arm assignment. Patients and staff are blind to study arm; staff are blind to patients' baseline or post-intervention assessments. Outcome measures are assesed after sessions 10, 20 and 30, and 12 weeks post-TCT. See "Research Strategy" for details.

General Statistical Approach: Hypothesis tests are 2-sided (α=.05). Type I error is controlled by selecting optimal *a priori* targets for primary analyses; secondary analyses use false discovery rate corrections. When appropriate, missing data are handled via multiple imputation or full information maximum likelihood estimation. Expected drop-out rates and ways to minimize them are described in the parent application.

Differential drop-out rates from PBO vs. AMPH arms are monitored. Clinical follow-up is pursued with all patients exiting the study through post-TCT week 12.

Cohort differences: For between-subject contrasts (drug), baseline demographic, clinical, antipsychotic/anticholinergic load and performance differences are tested and used as covariates if indicated. Sex differences are tested for all measures.

Procedures for Alternative or Unexpected Results: Linear mixed-effects models allow for the analysis of all available data using an intent-to-treat approach. This approach, however, assumes that data are missing at random. A potential concern is nonrandom dropout, particularly when dropout is related to the (missing) outcomes; that is, not missing at random. To make the missing at random assumption more plausible, we will explore the use of auxiliary variables. Auxiliary variables are covariates that predict both dropout and outcomes. By adding auxiliary variables to our models, nonrandom dropout is corrected by conditioning treatment effects on the predictors of dropout.

Primary Efficacy Endpoint(s): Aim 1: Target engagement – APS learning will be significantly greater under AMPH vs. PBO conditions. Aim 2: Pilot trial - Compared to PBO arm patients, AMPH arm patients will exhibit greater clinical benefit, as indicated by one or more of the following: greater, faster or more durable improvements in PANSSt (reduced scores; see below, "Go/No-Go"), MCCBc (increased scores) or WHODAS (reduced scores).

Secondary Efficacy Endpoint(s): Aim 3: Biomarkers - baseline levels of specific neurocognitive or EEG-based measures, or changes in specific measures with initial AMPH exposure, predict greater sensitivity to AMPH-enhanced neurocognitive, clinical or functional gains over 30 sessions of TCT.

Analyses testing primary and secondary hypotheses will include all subjects' clinical, neurocognitive and functional assessments. Data from subjects who do not complete the full study are carried forward in an "intent to treat" design, and their data are analyzed using linear mixed models. Separate analyses will compare completion / drop-out rates across arms, as well as a 14-item 7-point Likert scale assessing treatment satisfaction.

Go/No-Go decision and primary and secondary endpoints

"Go/No-Go" decisions are based on traditional statistical significance (Aim 1) and effect sizes (Aim 2). Formal statistical testing is conducted using linear mixed-effects (LME) models; hypothesis tests are 2-sided (α=.05). Model parameters are estimated via the R lme4 package and Cohen's d is estimated via the EMAtools package. Type I errors are minimized by constraining the number of primary analyses; secondary analyses use false discovery rate corrections.

PANSSt scores will be assessed at specific intervals. **It is important to emphasize that the anticipated symptomatic gains will reflect an enhanced impact of TCT, rather than a direct therapeutic impact of AMPH (5 mg, 2-3 times per week).** Because TCT is associated with improvements in both positive and negative symptoms, our *a priori* hypothesis is that a PACT effect will be evident in both positive and negative symptoms. Hence, PANSSt is used as the primary metric of symptomatic gains. A greater reduction in PANSSt score will be detected using data from baseline through TCT 30h, by a significant main effect of group (arm) or a significant interaction of group x session, with post-hoc contrasts revealing significant group differences at TCT 30h. A faster reduction in PANSSt scores will be detected using these same data by a significant interaction of group x session and a significant group difference at a point prior to TCT 30h. A more durable reduction in PANSSt scores will be detected using data from TCT 30h through 12 weeks post-TCT, by a significant main effect of arm or a significant interaction of group x post-TCT day, with post-hoc contrasts revealing significant group differences 12 weeks post-TCT.

Comparable analytic approaches will be used for neurocognition (MCCBc) and function (WHODAS) as described in "Research Strategy". Data will be described as means with SEM. Tests for normalcy will be applied to all data and non-parametric comparisons will be used if normalcy is not demonstrated. The Analysis Set for these data will be all subjects who have completed 30h of TCT and post-TCT clinical, neurocognitive and function assessments. When appropriate, missing data are handled via multiple imputation or full information maximum likelihood estimation.

Secondary regression analyses assess the predictive value of putative biomarkers on clinical, neurocognitive or functional gains from the addition of AMPH to TCT. Future studies with larger samples will assess potentially more informative models, e.g. Path Analysis to study moderating effects of APS AMPH sensitivity and biomarkers on both indirect (moderated-mediation) and direct paths between treatment and outcome.

Power analyses

Consistent with the FOA, this study will adequately power tests of target engagement (Aim 1) for traditional significance, while the "Go" signal for therapeutic impact (Aim 2) is based on effect size (d=0.5). To detect target engagement with d=0.5, 80% power, alpha=0.05, test-retest correlation =0.8, and a linear treatment effect, n=27 randomized participants are required per group. This is a very conservative estimate of d for target engagement, since empirically, d for target engagement with 5 mg amphetamine was 0.85 (Fig. 3A). A future Confirmatory Efficacy trial will be powered to apply more robust approaches to Aim 2, e.g. Path Analysis[108] to study moderating effects of APS AMPH sensitivity and biomarkers on both indirect (moderated-mediation) and direct paths between treatment and outcome.

Project Modification: Statistical Analysis Plan

The original study design included separate, parallel assessments of amphetamine (AMPH) (vs. placebo (PBO)) and memantine (MEM) (vs. PBO) effects on TCT-associated clinical, neurocognitive and functional gains. However, these designs pre-dated the pandemic, and between the date that they were proposed (10/2019) and the project start date (8/2020), our institution implemented a pandemic-related shut-down of clinical research activities. Even with the gradual resumption of research activities, recruitment lagged significantly behind target levels. As a result, we made several pragmatic adjustments to the original study designs: 1) Two separate but parallel protocols were approved by the UCSD IRB (protocols #191811, #201502) and posted on ClinicalTrials.gov. Due to the low recruitment numbers, the two parallel studies were combined and placebo groups were pooled, to yield a 3-group design 2) subject randomization changed from 1:1 to a 2:1 active : placebo ratio; and 3) adjustments were made to the planned statistical analyses, appropriate to the reduced sample size.

Based on known benefits of TCT in psychosis patients, no "placebo TCT" condition (e.g., video game) was included, for ethical and pragmatic reasons. Since all subjects received TCT, clinical gains over time in all groups might reflect, to some degree, the benefits of TCT. Pharmacologic augmentation by AMPH would only be evident if/when benefits in "TCT+active drug" groups exceeded those in "TCT+PBO" groups.

The primary goal of this study was to assess the clinical, neurocognitive and functional gains associated with the addition of AMPH to a full course of TCT. In total, there were 3 primary outcome measures and 5 secondary outcome measures. The primary clinical outcome measure was the Positive and Negative Symptom Scale (PANSS) total score; positive and negative subscale scores were secondary outcomes. Because they were highly correlated with total scores (r=0.92 at T_1), PANSS general scores are reported but are not outcomes. The primary neurocognitive outcome was the MCCB Composite Score. The primary functional outcome was the World Health Organization Disability Assessment Schedule (WHODAS 2.0, 12-item). Secondary clinical outcome measures also included the Patient Health Questionnaire (PHQ-9), the Young Mania Rating Scale (YMRS) and the Psychotic Symptoms Rating Scale (PSYRATS) hallucinations subscale.

Clinical, neurocognitive and functional outcome scores were treated as continuous variables and analyzed in three ways. First, they were submitted to linear mixed-effects (LME) models[59] for both between-group (interaction model: TCT+AMPH vs. TCT+PBO) and within-group (time model: baseline vs. P30, all groups) analyses. LME analyses included data from all randomized subjects (including those who did not complete 30 AT sessions) and modeled random intercepts to account for within-subject dependencies. Second, because sample sizes fell well below target levels, to test the hypothesis that active drugs would augment the impact of a 30-session course of TCT, the effect size (Cohen's d) of the change from baseline to P30 ("difference score") was calculated for both between-group (TCT+AMPH vs. AT+PBO) and within-group (baseline vs. P30, all groups) comparisons. Between-group effect sizes were calculated as:

$$d = [(\text{drug score } T_1 - \text{drug score P30}) - (\text{PBO score } T_1 - \text{PBO score P30})] / SD_{wav}$$

where SD_{wav} is the weighted average standard deviation. Within-group effect sizes were calculated as:

$$d = (\text{score at } T_1 - \text{score at P30})/SD_{av})$$

where SD_{av} is the average of the standard deviations of the T_1 and P30 measures[60,61].

Third, to disentangle the salutary contributions of TCT alone vs. TCT+active drug to within-subject effect sizes, a "threshold" for positive active drug effects was set at "d=0.4 *above PBO values":*

d(active drug) $\geq$ d(PBO) **+ 0.4**).

Because clinical gains from TCT+PBO were expected to be approximately d=0.4, this added threshold of "d=0.4 above PBO" for drug-enhanced TCT meant that an estimated combined large effect size (at least d=0.8) would be required for a drug effect to be viewed as meaningful. Outcomes using thresholds from d=0.2 to d=0.6 were reported.

The use of effect size comparisons in addition to p-values is common in clinical trials research as a way to measure effects independent of sample size. It is particularly useful when, as in the current study, sample sizes fall below targets identified based on traditional power analyses. Effect size "thresholds" have been used by others to argue that TCT neurocognitive benefits exceed levels produced via PBO or practice effects (generally d$\approx$0.18).

Indirect indicators of clinical impact included subject attrition, C-SSRS responses, treatment satisfaction and adverse events (AEs). AMPH group subjects were evaluated for abnormal involuntary movements (AIMS) and for deleterious effects of AMPH cessation (AMPH Cessation Symptom Assessment). Analyses included non-parametric (Chi-Square) and parametric (rmANOVA) tests.

Linear mixed-effects models for outcome measures included all subjects who were randomized; df were estimated using the Satterthwaite approximation. Between- and within-group effect sizes for clinical, neurocognitive and functional measures were calculated based on all subjects who completed 30 TCT visits and P30 testing ("completers"). For all analyses, α was set at 0.05; for analyses of the 5 secondary outcome measures, findings were also evaluated with a more restrictive alpha level (0.01) to correct for multiple comparisons.