Metabolic Effects of Antipsychotics in Children (RO1MH072912)

NCT #: 00205699

Date: 3/28/18

Data Analysis Plan

Revised Final for JAMA Psychiatry and clinicaltrials.gov submission

Data management and quality control: The data management and quality control process will begin with the initial design of data collection forms (when needed) and data capture forms in SPSS (IBM SPSS; Armonk, NY). Quality control measures will include the flagging of out-of-range and inappropriately missing data items, the identification of forms that have not been completed for a particular subject, and confirmation that enrolled subjects satisfy eligibility criteria. To further ensure accurate and complete data, all data collection personnel will be asked to review forms as soon as they are filled out so that problems related to missing or illegible data items can be quickly corrected. Because forms will include the initials of the data collection person who completed the form, it will be easy to contact the appropriate individual when problems are identified.

Data analysis: Primary and secondary analyses for adiposity and insulin sensitivity outcomes were conducted using Intent-to-Treat (ITT) analyses (IBM SPSS; Armonk, NY) including randomized subjects. Primary outcomes were change in DEXA-measured adiposity (DEXA % total fat) and clamp-derived insulin sensitivity at muscle (percent change in Glucose Rd), with secondary outcomes of change in MRI measured adiposity and percent change in Glucose Ra and Glycerol Ra. Primary analysis for change in DEXA % fat used a likelihood-based mixed-effects model using time (0, 6 and 12 weeks) and medication group as independent variables, with Toeplitz covariance structure specified, based on Bayesian information criteria (BIC). The primary outcome analysis for insulin sensitivity, as well as the secondary outcome analyses for adiposity and insulin sensitivity, used repeated-measures analyses of covariance (ANCOVA) with baseline values of the dependent variables as the covariate (to address potential baseline influence on outcomes) and time and treatment condition as independent factors, testing for time by treatment condition as well as covariate interactions. When time by treatment interactions were significant, contrasts were used to test comparisons of interest; when not significant, treatment condition was removed from the model to calculate the main effect of time and any interactions. MRI abdominal fat models were run with compartment (subcutaneous versus visceral) as an additional 2-level factor, to test for time by compartment and time by treatment by compartment interactions. Exploratory analyses tested whether the effects of age, stimulant use, gender and race altered primary results, re-running primary analyses with an additional two- or three-level factor (e.g., yes/no stimulant); these exploratory analyses were corrected for multiple tests (Bonferroni; 0.05/4= 0.0125). Other exploratory analyses used ANCOVA as above (week 0 and 12) to support interpretation of primary/secondary measures (e.g., DEXA lean, clamped insulin concentration) or for clinical context (e.g., BMI %ile, psychiatric symptoms), with ANOVA used to test the effect of time within individual treatment groups. Effect sizes (Cohen's d) were calculated for primary and secondary outcomes.

Exploratory analyses include correlating the 12 week change in measures of insulin sensitivity and the corresponding change in percent fat with a view towards a more precise understanding of the degree to which changes in insulin sensitivity can be explained by changes in fat mass. Analyses of covariance that employ changes in insulin sensitivity as the dependent variable and include changes in percent fat as a predictor will provide estimates of the percent of variability explained (R^2) both by the change in fat mass alone and by the combination of this measure and the covariates already noted. These analyses will also provide information about the degree to which the covariates are independently predictive of changes in insulin sensitivity. Other exploratory analyses will involve the role of regulatory hormones such as cortisol, glucagon, and leptin. Since these hormones will be measured at baseline and at 6 and 12 weeks, mixed model repeated measures analysis of variance will evaluate the effect of treatment on these hormones and will provide information about the pattern of change during the 12 week follow up period. A final set of possible analyses will involve the use of logistic regression to determine factors that are associated with the presence of metabolic syndrome at baseline, to be performed only if baseline levels are meaningful. While we will also use logistic regression to explore factors that are predictive of the development of metabolic syndrome during follow-up and will evaluate between group differences in the incidence of syndrome development, we do not anticipate that large numbers of subjects who do not have this syndrome will develop it during the relatively brief 12-week follow-up. Thus, we expect that the statistical power associated with these latter analyses will be small and view them as exploratory hypothesis generating analyses.

In all of the analyses we perform, we will give careful attention to the appropriateness of the statistical procedures that are employed. Thus, all analyses will be preceded by graphical assessments of the data to evaluate the distribution of particular variables and to determine whether outliers are present. When we perform analyses of covariance, we will routinely evaluate regression residuals to ensure that the model is appropriate. Similarly, we will evaluate the Hosmer-Lemeshow goodness of fit statistic to assess the appropriateness of logistic regression models. When variables are poorly distributed and assumptions for a given analytic procedure are violated, we will explore and implement remedial measures that include data transformations and semi-parametric approaches based on the ranks of the data if satisfactory data transformations cannot be found. We have defined only two

primary specific aims with the primary comparison being baseline to 12 weeks both because of the importance of those aims and to minimize multiple comparisons concerns. Because of the multiple comparisons issue, p-values associated with the secondary and exploratory analyses we perform will be interpreted cautiously as hypothesis-generating rather than as hypothesis-confirming.

Sample size and statistical power: The target sample size is 80 subjects in each of the three study groups. Prior data from young adults (i.e., NIMH and industry funded first episode studies where change in % body fat may be crudely estimated from known change in body weight and the assumption that increases occur in body fat rather than lean muscle), which may underestimate changes in children, suggests that percent body fat increases of 10 $\pm$ 4% may occur in the olanzapine group, 5 $\pm$ 4% in the risperidone group and 0 $\pm$ 4% in the aripiprazole group (e.g., 10 lbs fat increase in a 70 lb subject who starts at 20% body fat yields 30% body fat). Based on this, the power for a two-sided test at the 0.05 level of significance to compare any one group with any other group will be essentially 1. This substantial power will permit us to test hypotheses about changes in percent body fat in important subgroups that are defined by age, gender, race, and whether the subject is or is not taking stimulants (e.g., planned younger off-stimulant subgroup sample size is 20), as well as explore other predictors. Because we do not have reliable prior data on likely pooled baseline to endpoint or between-group differences in insulin sensitivity measures like the glucose disappearance rate, we base our power computations on the effect size that is detectable with 80 subjects per group. Using a two-sided test at the 0.05 level of significance, the target sample size implies that we can detect a between group difference in the mean disappearance rate equal to half of a standard deviation with a power of 0.88.

A secondary interest of the proposed research will be whether the younger prepubescent children between the age of 6 and 11 respond differently from children between the ages of 12 and 18. In our existing clinic system, a sample of 527 children who had the study eligible diagnostic codes indicated that 69.3% were in the younger age group. Even assuming a worst-case scenario that we fail in our planned efforts at over-recruitment to the older age groups and that these percentages continue as we recruit for the planned study, we anticipate that in each study arm, we will have about 56 subjects in the younger age group and 24 in the older age group. With these numbers and using within group analyses, we will have a power of 0.81 to detect a response rate difference in an outcome measure in younger as compared to older individuals that is 70% of a standard deviation, with the power increasing to 0.90 if the effect size is 80% of a standard deviation. If there are no statistical interactions between treatment group and age category, it will be reasonable to combine all three groups in asking about age-related differences so that worst case sample sizes will be 168 in the older group and 72 in the younger group. This dramatically increases statistical power to the extent that we will be able to detect an effect size equal to 40% of a standard deviation with probability of 0.81. The power increases to 0.94 under this no interaction assumption if the effect size comparing age differences is 50% of a standard deviation.