Statistical Analysis Plan

for

Study Protocol

Intervention and Outcomes in Duarte Galactosemia

Principal Investigator Judith Fridovich-Keil

IRB00081271

NCT02519504

Date 04/01/2019

Statistical Analysis Plan:

Our Statistical Analysis Plan, prepared in follow-up to multiple meetings of a "Data Analysis Subgroup" from our research team, is described below. In short, data were cleaned and relevant covariates were identified for each outcome. Next, our cohort of 350 volunteers (206 cases and 144 controls) was randomly subdivided into independent "discovery" (n=87) and "validation" (n=263) sets that were approximately balanced with regard to case/control composition and covariates. We then tested each of the 73 outcome variables collected for each subject for possible association with case/control status using data from the discovery cohort; though none met the bar for statistical significance after multiple test correction we simply ranked them by ascending p value. Finally, we tested the 10 outcomes with the smallest p values from this ranking for (a) possible association with case/control status in the validation set, and also (b) possible association among cases in the validation set with milk exposure in the first year of life. All statistical analyses were performed in R.

How the data were cleaned:

1. We imported study data from REDCap (PCORIDGStudy_R_2017-XX-XX_XXXX.r) into R.
2. We filtered subjects by participation status to include only those who completed both Part 1 and Part 2.
3. We filtered out subjects marked for exclusion from analysis (e.g. after testing we realized they should have been excluded).
4. We filtered out subjects whose DG status was unknown.
5. We created a new variable representing the highest education level of either parent/guardian.
6. We created a new ordinal variable for breast milk exposure.
    - Value (0, 1, 2, or 3) indicating the number of time intervals (birth to 1 month, 2 to 6 months, 7 to 12 months) during which the subject drank at least some breast milk.
7. We created new variables for average ABERS latency values at each wave (i.e., average “aber_w1_lat” and “aber_st_w1_lat”).
8. For each variable in the dataset we:

a. Indicated category label: administrative variable, covariate, DG status variable, dairy exposure variable, or outcome variable.
b. For covariates, we indicated whether the variable should be considered in Aim 1 (DG/non-DG analysis), Aim 2 (DG-only diet analysis), or both.
c. For outcome variables, we indicated whether the variable should be considered in Aim 1 (DG/non-DG analysis), Aim 2 (DG-only diet analysis), or both.
d. For outcome variables, we indicated the general domain of development under which it belonged.
e. For each outcome variable, we indicated whether it had already been adjusted to account for any covariates and, if so, which ones (e.g., assessment score already normed for age and gender).
f. Indicated data source (Part 1 survey, Part 2 testing).
g. Indicated variable type (continuous, ordinal, categorical, binary).

9. For each variable, we plotted all data points and looked for outliers.
   a. For continuous variables, we used a threshold of ±3 SD from the mean.
   b. If found, we confirmed with the testing team whether the outlier was genuine or a scoring/data entry error.
10. We cleaned and aggregated sparse categorical covariate levels where needed.
11. We compared "missingness" among outcome measures between DG cases and controls.
    a. We tested whether the proportion of DG cases missing data on a given outcome differed significantly from the proportion of controls missing that outcome data.
    b. We used a chi-square test if the number missing was large (i.e., expected counts >5) or Fisher's exact test if the number missing was small (i.e., expected counts <5).
    c. If the proportion of missing data was significantly different between cases and controls, we could either: (1) exclude that outcome from analysis, or (2) impute missing data conditional on phenotype [see *Statistical Analysis with Missing Data* by Little and Rubin].
12. We calculated summary statistics for variables (from both Part 1 and Part 2 sources) across: (1) all subjects, (2) non-DG subjects only, (3) DG subjects only, (4) dairy-exposed DG subjects only, and (5) non-dairy exposed DG subjects only.

- We used mean, SD, median, and IQR for continuous variables.

13. For continuous variables, we checked to see if they followed a normal distribution using the Shapiro-Wilk Normality test. For non-normally distributed continuous variables, we tried transformation to achieve normality. If a log or square-root transformation failed to work, we used an inverse-normal transformation.

Covariate selection:

As we tested multiple outcome measures for association with DG status (Aim 1) and for association with diet among DG cases (Aim 2), for each of the outcomes considered, we identified the relevant covariates to adjust for. Using the steps described below, we determined the subset of covariates to be included in subsequent models for each outcome.

1. Using the information provided in step #8 above, we identified all variables that were to be initially considered as potential confounders for each outcome measure.
2. With these variables as predictors and the outcome measure of interest as the response variable, we used the R package *glmnet* to fit a generalized linear model with lasso regularization.
   a. We used 10-fold cross-validation to determine the optimal lambda value.
   b. We identified the subset of variables with non-zero coefficients and flagged these as covariates for subsequent models of the outcome.
   c. Please note: This approach assumed observations were independent and did not take into account the shared familial structure of the dataset. As described below, in later analyses we used a mixed-model approach to account for relatedness among family members in the study.

Dividing our full cohort into independent discovery and validation sets:

Although data on 73 outcome variables were collected, our study was only adequately powered to test ~20 hypotheses due to the need to correct for multiple testing. As we sought both to test phenotypic differences between DG cases and controls, and also to test phenotypic differences associated with diet among cases, this substantially reduced the number of phenotypes we could consider without compromising the power of the data set.

In order to find a balance between our desire to maintain adequate statistical power and our obligation to find differences in the data wherever they might exist, we first divided our sample into independent "discovery" and "validation" sets. We used the discovery set to identify the 10 outcome measures of greatest interest (smallest p value). We then used this subset of outcomes for a focused analysis of the validation set at an appropriate Bonferroni-adjusted significance threshold.

1. We randomly selected ~25% of Part 2 study subjects for assignment to the discovery set.
    - To ensure independence of the discovery and validation sets, siblings were always assigned to the same set.
2. We assigned the remaining 75% of study subjects to the validation set.
3. We repeated the random assignment process until an appropriate mix of DG cases/controls and dairy-exposed/non-exposed cases (close to original distribution in the full dataset) was achieved in each set.

Using the discovery data set to rank order outcome variables to test in the validation data set:

1. In the discovery set, we fitted a mixed-effect regression model (linear, logistic, or ordinal) for each outcome variable that included DG status and all necessary covariates as predictors.
2. The mixed model approach allowed for clustering at the family level and therefore took into account relatedness among study subjects.
3. We ranked the models by resulting p-value for the DG status regression coefficient, and selected the 10 top-scoring outcomes between cases and controls.
4. Even if no models showed significance at raw $p<0.05$ for association with DG, we choose the 10 outcomes with the smallest p-values.

Aim 1: Identify phenotypes associated with DG

1. We fit comparable models for the 10 outcomes showing the smallest p values in the

discovery set using data from the independent validation set.

2. Our goal was to identify any outcomes whose models showed significant association with DG status at the Bonferroni-adjusted p-value threshold ($\alpha = 0.05/20 = 0.0025$) in the validation set. There were none.

Aim 2: Identify phenotypes associated with diet among DG cases

While the random assignment process described above yielded a discovery set with a satisfactory number of both cases and controls to assume it was representative of the full study population, the number of dairy-exposed cases and non-exposed cases in the discovery set was not large enough to confidently consider it a representative sample. For this reason, we tested the same 10 outcome variables as above for association with diet among DG cases.

1. We filtered the discovery set generated as described above to include only DG cases.
2. Using the discovery set, we fitted a mixed-effect regression model (linear, logistic, or ordinal) for each of the 10 selected outcome variables that included milk exposure and all necessary covariates as predictors.
3. Our goal was to identify any outcomes whose models showed significant association with diet among cases at a Bonferroni-adjusted p-value threshold ($\alpha = 0.05/20 = 0.0025$) in the discovery set. There were none.