Statistical Report for Treatment Evaluation

Treatment Evaluation Statistical Tool, TEST 1

07/27/2022

Table of Contents

# 1 TEST Session Summary

|  |  |
| --- | --- |
| Statistician | Cheng Su |
| Sessionmode | Confirmatory |
| Name |  |
| Studyid | TB21-02 |
| Email |  |
| Program | BMN000 |
| Department | Dep 1 |
| Project | pan\_program\_non\_project\_specific\_support |
| Uuid | e3e93104-48e2-4cf0-922a-9e905a467a8d |
| Base Dir | C:/Users/Christopher.Wilson/Documents/bmrn-test |
| Rel Path Home | test\_output/BMN000/pan\_program\_non\_project\_specific\_support/TB21-02/e3e93104-48e2-4cf0-922a-9e905a467a8d |
| Full Path Home | C:/Users/Christopher.Wilson/Documents/bmrn-test/test\_output/BMN000/pan\_program\_non\_project\_specific\_support/TB21-02/e3e93104-48e2-4cf0-922a-9e905a467a8d |
| Full Path Files | C:/Users/Christopher.Wilson/Documents/bmrn-test/test\_output/BMN000/pan\_program\_non\_project\_specific\_support/TB21-02/e3e93104-48e2-4cf0-922a-9e905a467a8d/files |
| Timestamp | 2022-07-27 18:23:22 |

# 2 Study Background and Objectives

# 3 Statistical Methods

## 3.1 Overview

In a typical study supported by TEST 1, data is collected over time, and the primary interest is to evaluate a treatment in multiple doses. The study typically includes two types of subjects, wild type and non-wild type, and several controls, such as vehicle, negative control, positive control and other comparators. We will denote the Wild Type animals as the Wild Type group and the non-Wild Type groups as vehicle, negative control, positive control, other comparators, and the treatments as doses. Based on the study design and known study objective, TEST 1 defines a set of group comparisons to be evaluated statistically.

TEST 1 implements an algorithm that carries out a multiple-step analysis for statistical comparisons of groups. The statistical comparisons are conducted through mixed effect repeated measure modeling. TEST 1 is designed to automatically determine the best data transformation, check model assumptions (such as normality and homogeneous variance) and the appropriate correlation structure for the repeated measures (data across time points). In addition, TEST 1 generates linear contrasts and carries out multiple testing adjustment for the pre-defined tests.

Explanations of these key steps are described below, and the workflow of TEST 1 is described in appendix (section 6).

## 3.2 Basic Model

We define the basic model as a model that only considers the vehicle and treatment groups. These groups will be used to determine if a transformation is needed and which transformation is most appropriate. If the normality and homogeneous variance assumptions are not met in the basic model then further assistance is required and the consulting statistician will be contacted. The treatment and vehicle groups are used to remove the risk of contamination from the controls or wild type groups.

## 3.3 Mixed Effects Modeling

TEST 1 uses mixed effects models to evaluate statistical differences between multiple experimental (treatment) groups across multiple times points. The model evaluates factors including treatment, time, and the interaction between treatment and time. In addition, TEST 1 determines the appropriate correlation structure to account for the correlation between different time points. Some common correlation structures in consideration include auto-regressive (AR1) correlation structure which assumes that the correlation between time points decays at an exponential rate, compound symmetry (CS) correlation structure assumes that the correlation between time points is constant between any given two time points, and unstructured is the most flexible and has no constraints.

In TEST 1, we also allow mixed effects models can use different variance/covariance structure for each group separately as needed.

## 3.4 Normality Assumption and Transformation

A Shapiro-Wilk test is conducted to determine if the residuals from a linear model with a treatment, time, and treatment and time interaction term are normally distributed. The Shapiro Wilks test assumes that the residuals are normally distributed and looks for evidence for non-normality. If the Shapiro-Wilk test is rejected (there is enough evidence that the residuals do not follow a normal distribution), then a Box-Cox transformation is conducted to suggest an appropriate transformation. Note that the Shapiro Wilks test is rejected if the p-value is less than 0.05. Then the Shapiro-Wilk test is repeated on the residuals of the transformation to confirm that the residuals of the transformed data follow a normal distribution. If the Shapiro-Wilk test is rejected, then we recommend further discussion with a statistician as a transformation did not make data follow the normality assumption.

In experiments that have technical replication, all of the data will be used to check the normality assumption and subsequently determine the most appropriate Box-Cox transformation (treating technical replicates as independent). After this step, the technical replicates will be averaged at each time point for each subject. The resulting dataset will be used for all of the remaining analysis steps.

## 3.5 Checking for Similar Variance between Groups

### 3.5.1 Basic Model

For this application, we require that the variance for each of the groups in the basic model to be similar. To verify this assumption, first the variance is determined for each group and at each time point, and then averaged across the time points. In addition, a likelihood ratio test (LRT) is conducted between a model that estimates one common variance and a model that estimates an individual variance for each group. In this case, LRT helps determines if a more complex model fits the data significantly better than a less complex model. Thus, a significant (p-value < 0.05) LRT provides evidence that it would be beneficial to model the variances of the groups separately. The variance between groups will be considered different if one the following two conditions are met:

1. If the p-value from the LRT is less than 0.05 (statistically significant) and the fold change between any group and the common variance is greater than 2, or
2. If the p-value from the LRT is greater than 0.05 (not statistically significant) and the fold change between any group and the common variance is greater than 3.

If the variance between groups are determined to be different, then no further analysis will be conducted and the consulting statistician will be contacted. Otherwise, the application will move forward to the next modeling step.

### 3.5.2 Controls and Wild Type

Once the variance for groups within basic model is determined to be similar, then a similar procedure is followed as above. A LRT is conducted for a model that has one common variance and a model with a common variance for the groups in the basic model and a different variance for the controls and wild type. A similar set of criteria as above to determine if a group has a different variance than the basic model:

1. If the p-value from the LRT is less than 0.05 and the fold change between any group and the of the basic model is greater than 2, or
2. If the p-value from the LRT is greater than 0.05 and the fold change between any group and variance of the basic model is greater than 3.

If a group is determined to have a difference variance than the basic model then a estimation of the variance/covariance will be separate from the basic model. Otherwise, the application will move forward to fit the final mixed model with one single common variance component.

## 3.6 Selection of Correlation Sturcture

The Akaike Information Criterion (AIC) is used to select the covariance structure that is most appropriate. The AIC strikes a balance between model complexity (number of parameters) and quality of the model fit (LRT). The candidate covariance structures are AR1 (ARH1), CS (CSH), Toeplitz (TOEP), and unstructured (UN) where the AR1 and CS are simplistic structures (only estimates one additional parameter) while UN correlation structure is the most complex (requires the estimation of additional parameters where is the number of time points). AR1H and CSH covariance structure allows for each time point to have a different variance.

## 3.7 Comparison between Experimental Groups

In this type of studies, there could be nine group comparisons:

| Label | Comparison | Purpose of Comparison |
| --- | --- | --- |
| A | Wild Type vs. Vehicle | Verify Disease Model. |
| B | Positive Control vs. Vehicle | Verify the experiment. |
| C | Wild Type vs. Each Dose | Can any doses reverse the disease? |
| D | Vehicle vs. Each Dose | Are any treatments effective? |
| E | Wild Type vs. Positive Control | Does Positive Control reverse disease? |
| F | Positive Control vs. Each Dose | Which doses are similar to Positive Control? |
| G | Each Dose vs. Every Other Dose | Do doses differ from each other? |
| H | Negative Control vs. Vehicle | Rule out matrix effect. |
| I | Negative Control vs. Each Dose | If H is rejected, then show Negative Control is not as good as treatment. |

## 3.8 Multiple Hypothesis Testing

Multiple comparisons arise when a statistical analysis involves multiple simultaneous statistical tests, each of which has a potential to produce a discovery. A stated confidence level generally applies only to each test considered individually, but often it is desirable to have a confidence level for the whole family of simultaneous tests. Controlling the overall Type 1 error rate, as opposed to the Type 1 error rate of each individual test, allows us to avoid rejecting a comparison simply due to random error. For instance, if 100 comparisons were conducted each with 5% Type I error rate, we would expect 5 results to be false positives. P-value adjustment allows us to control the probability of obtaining at least one false positive in a family of tests.

For each one the comparison groups described above, linear contrasts are constructed to allow for simultaneous testing of the groups, as wells as testing the groups at the pre-selected time point and over time course average. To adjust for multiple comparison and preserve family wise Type I error rate, a simulation based method is used to determine the critical values and p-values based on sampling from the multivariate t distribution.([Liu et al. 2007](#ref-Liu)) ([Edwards and Berry 1987](#ref-berry))

For comparisons between doses, and doses and control there are multiple tests. To adjust for multiple comparison and preserve family wise Type I error rate, a simulation based method is used to determine the critical values and p-values based on sampling from the multivariate t distribution.([Liu et al. 2007](#ref-Liu)) ([Edwards and Berry 1987](#ref-berry)) There will be no p-value adjustments for the comparision between controls.

## 3.9 Change from Baseline Analysis

The user will be allowed to select whether the raw/transformed values are to be analysed or to analysis the change of each observation from a subject’s baseline measurement. The normality check and Box-Cox transformation will be conducted ignoring the baseline, then the change from baseline will be the response studied for the remainder of the analysis.

## 3.10 Software

TEST 1 was developed using the R statistical software ([R Core Team 2016](#ref-Rbase)). The nlme::gls ([Pinheiro et al. 2013](#ref-nlme)) (this syntax means) [function library]::[specific function]) was used to construct the mixed effects models, and then the emmeans::emmeans function ([Lenth 2022](#ref-emmeans)) was used to compute the group and pairwise comparison summary statistics, as well as, adjust the p-values for multiple comparisons. R Shiny ([RStudio, Inc 2013](#ref-shiny)) was used to generate the web tool and this reported was generated using RMarkdown ([Allaire et al. 2020](#ref-rmarkdown)). The interactive plots within the web application were generated using the plotly::ggplotly function which converts a plot from ggplot2 ([Wickham 2016](#ref-ggplot2)) to plotly ([Sievert 2020](#ref-plotly)). The prism-style plots were generated using the ggprism([Dawson 2021](#ref-ggprism)) package which is an extension to the ggplot2 package.

# 4 Results

## 4.1 Assumptions and Correlation Structure Results

A Square Root was applied to the data. of the data was determined by the algorithm.

An AR1 was determined by the algorithm.

Wild Type had a different variance from the basic model.

## 4.2 Graphical Summary of Residuals Before and After Transformation

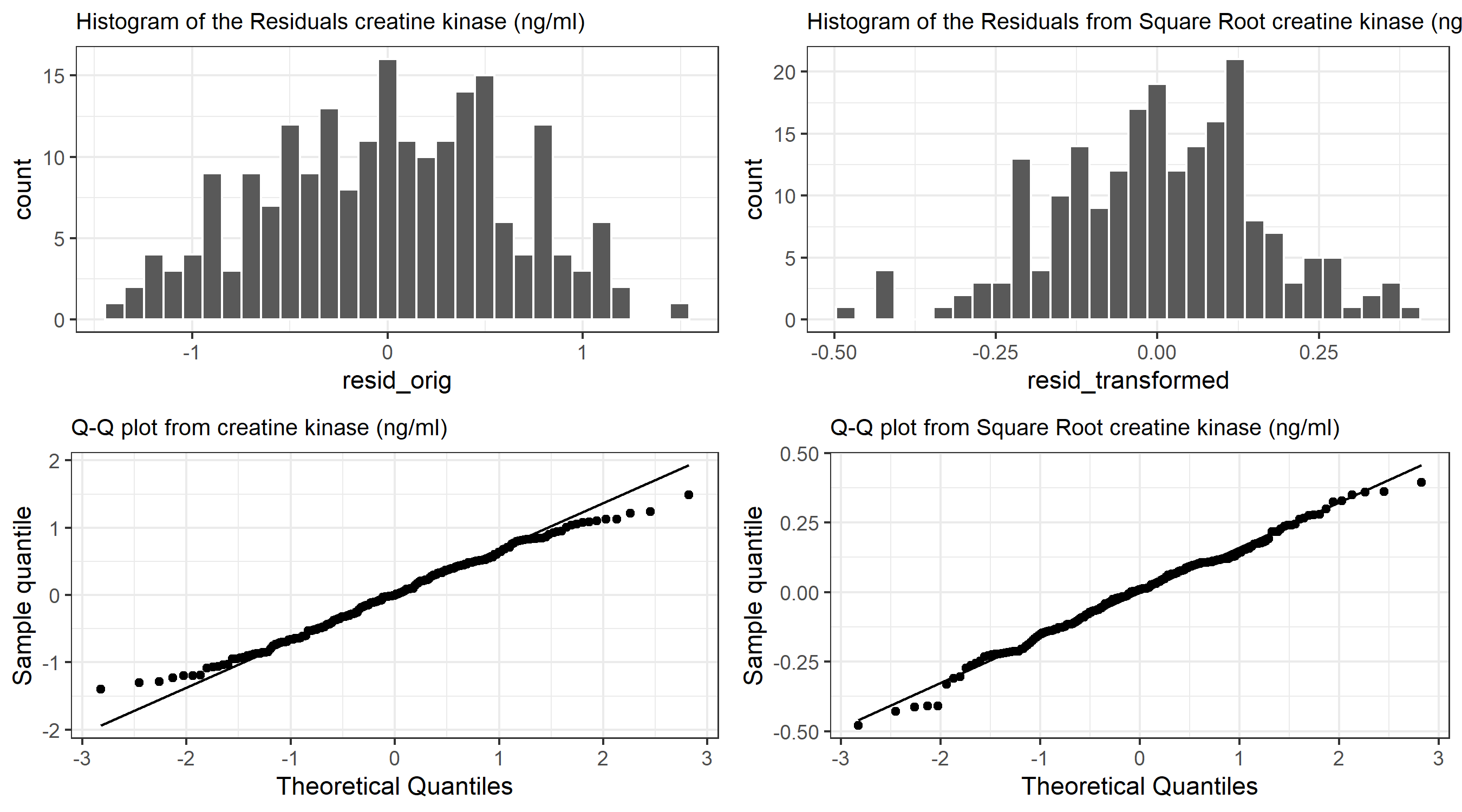


Figure 4.1: Histogram (top) and Q-Q plots (bottom) to assess the normality of the residuals

## 4.3 Plots

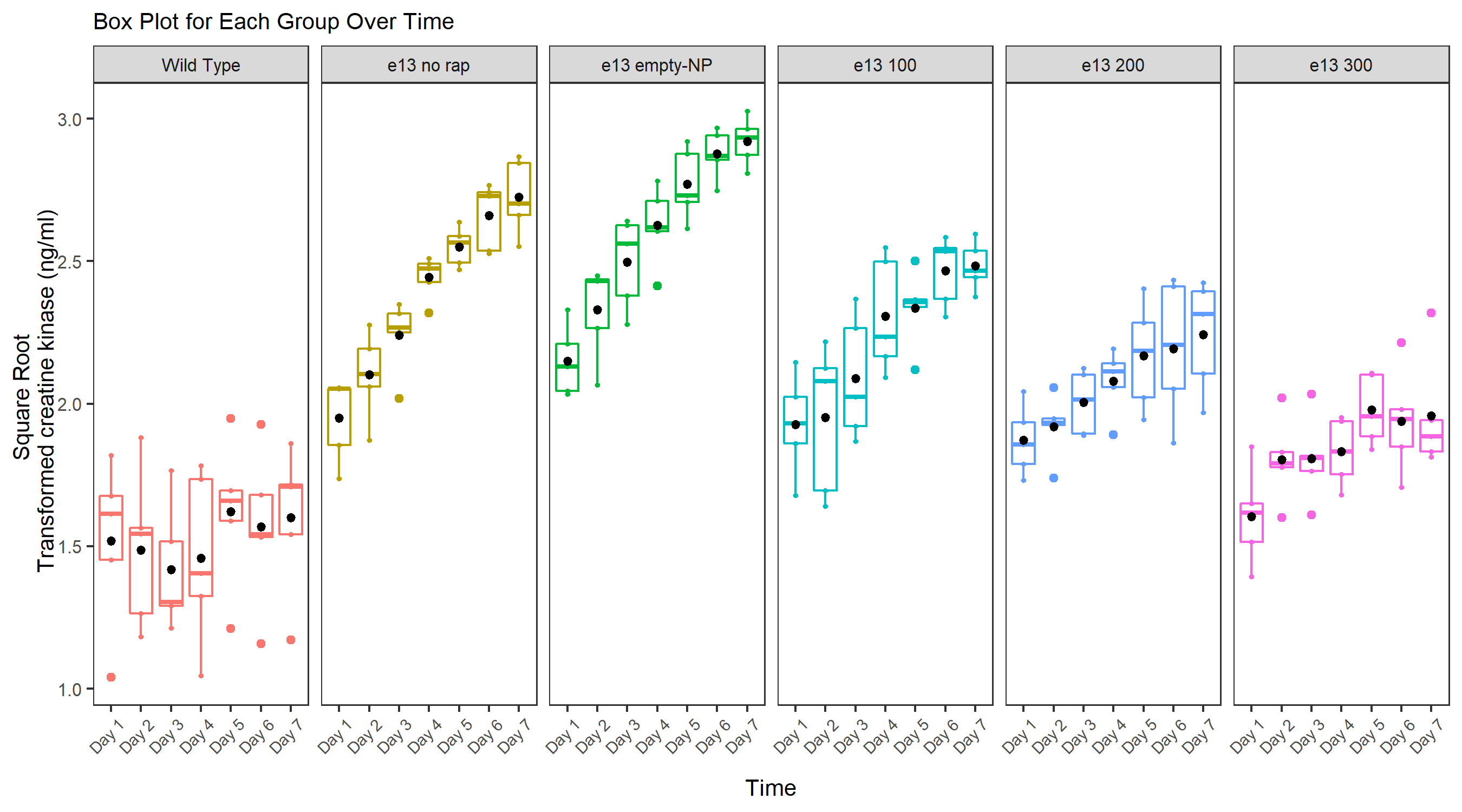


Figure 4.2: Box plot for each group and time.

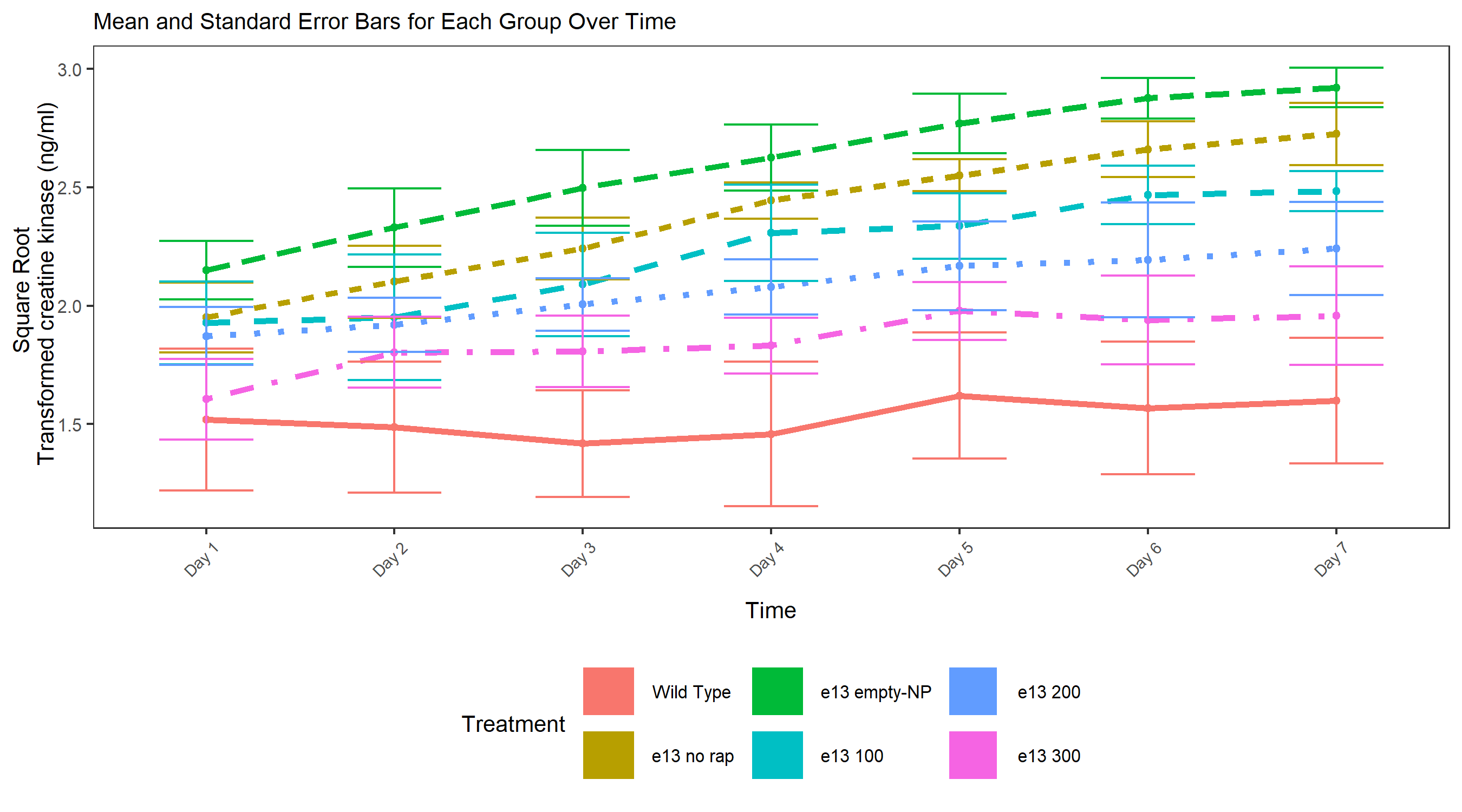


Figure 4.3: Group level trajectory across time where the line corresponds to the group average of Transformed Scale. The vertical bars are standard error bars for each group at each time point.

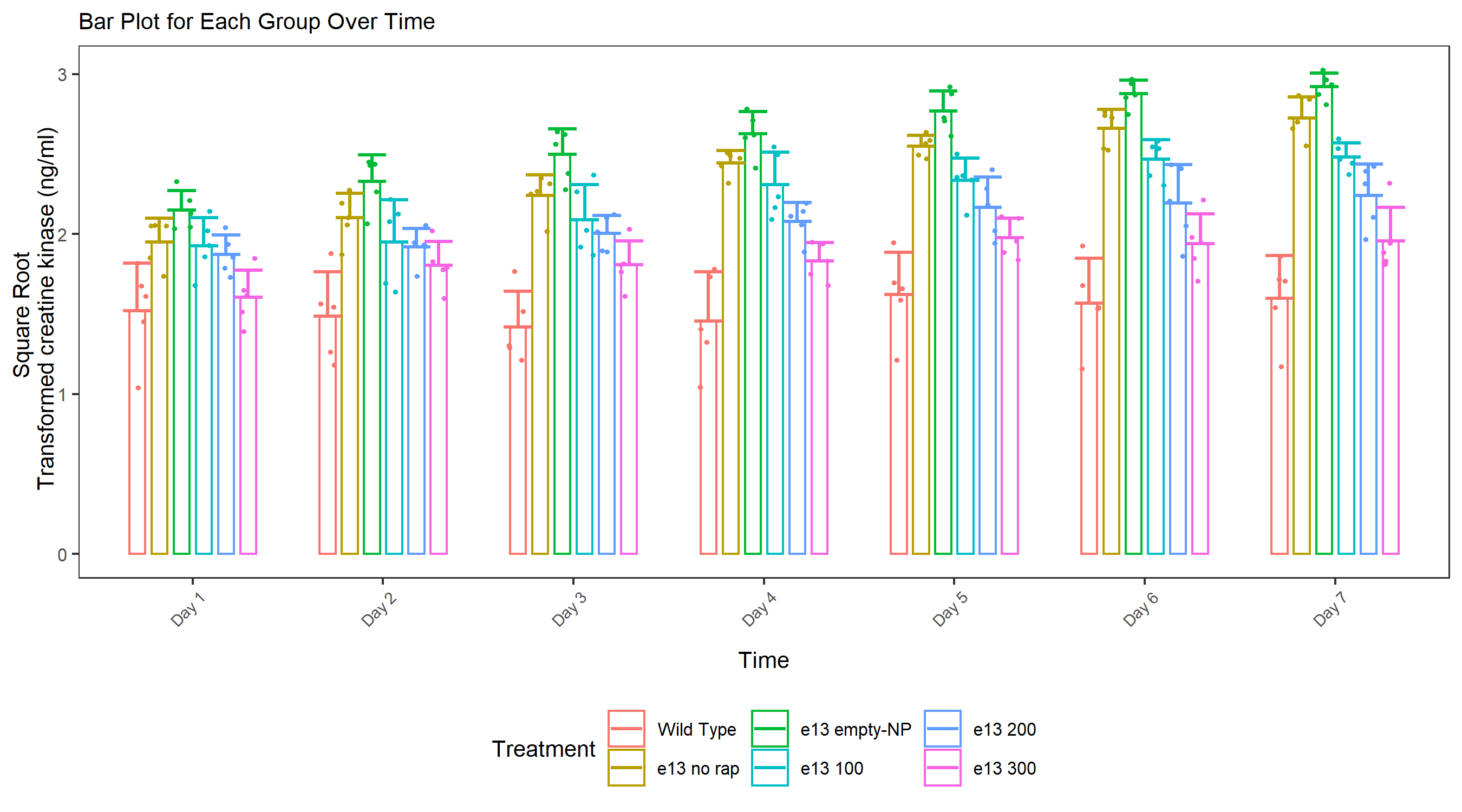


Figure 4.4: Bar plot for each group and time.

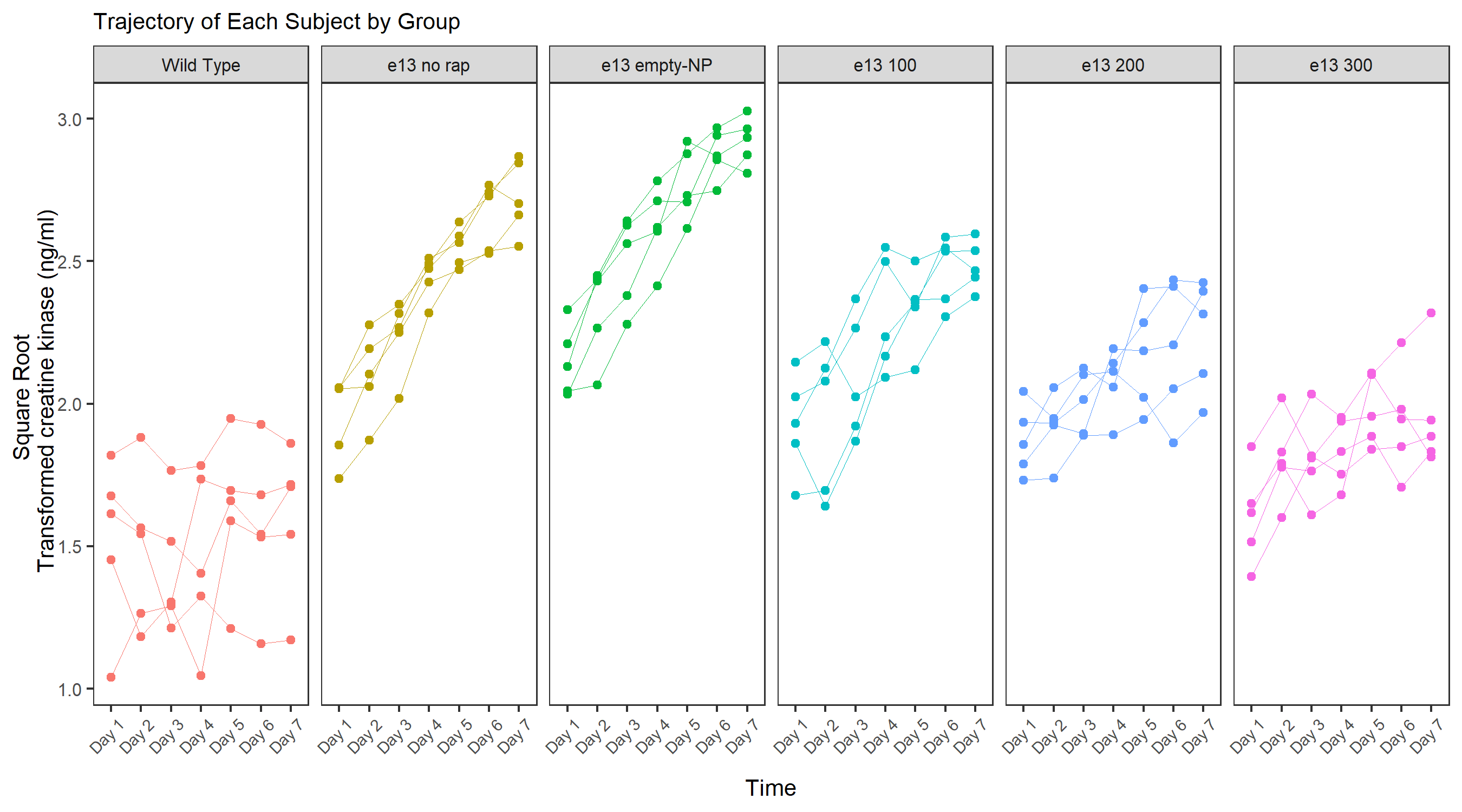


Figure 4.5: Subject level trajectory across time where the line corresponds to the average of Transformed Scale response in the presence of techinical replicates.

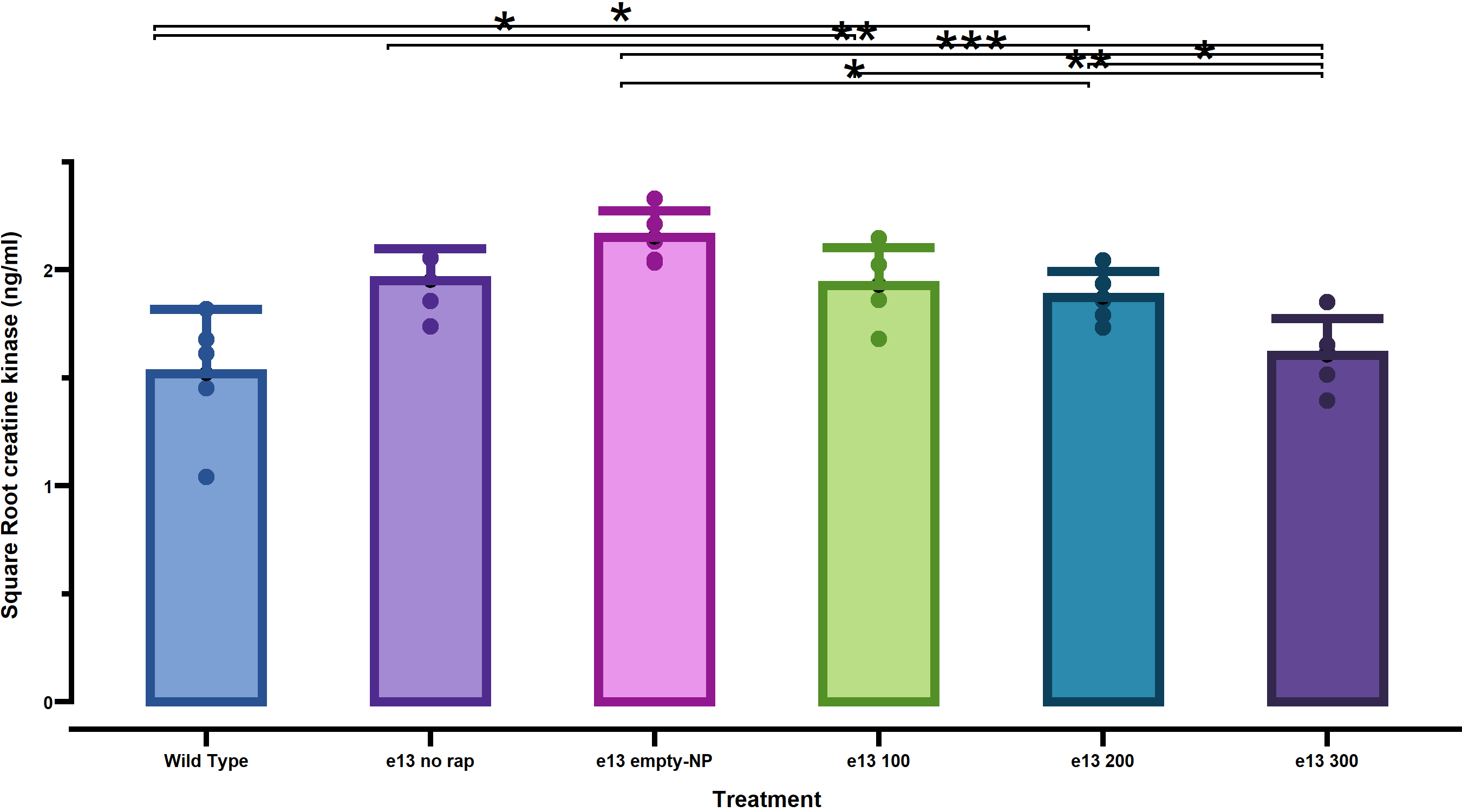


Figure 4.6: Boxplot of Transformed Scale response values at Day 1. The stars above the horizontal bars correspond to the magnitude of the p-value comparing two groups where “\*,” “\*\*,” and “\*\*\*” represent 0.01 <= p-value < 0.05, 0.001 <= p-value < 0.01, and p-value < 0.001, respectively. A p-value summary will only be shown form p-value less than 0.05.

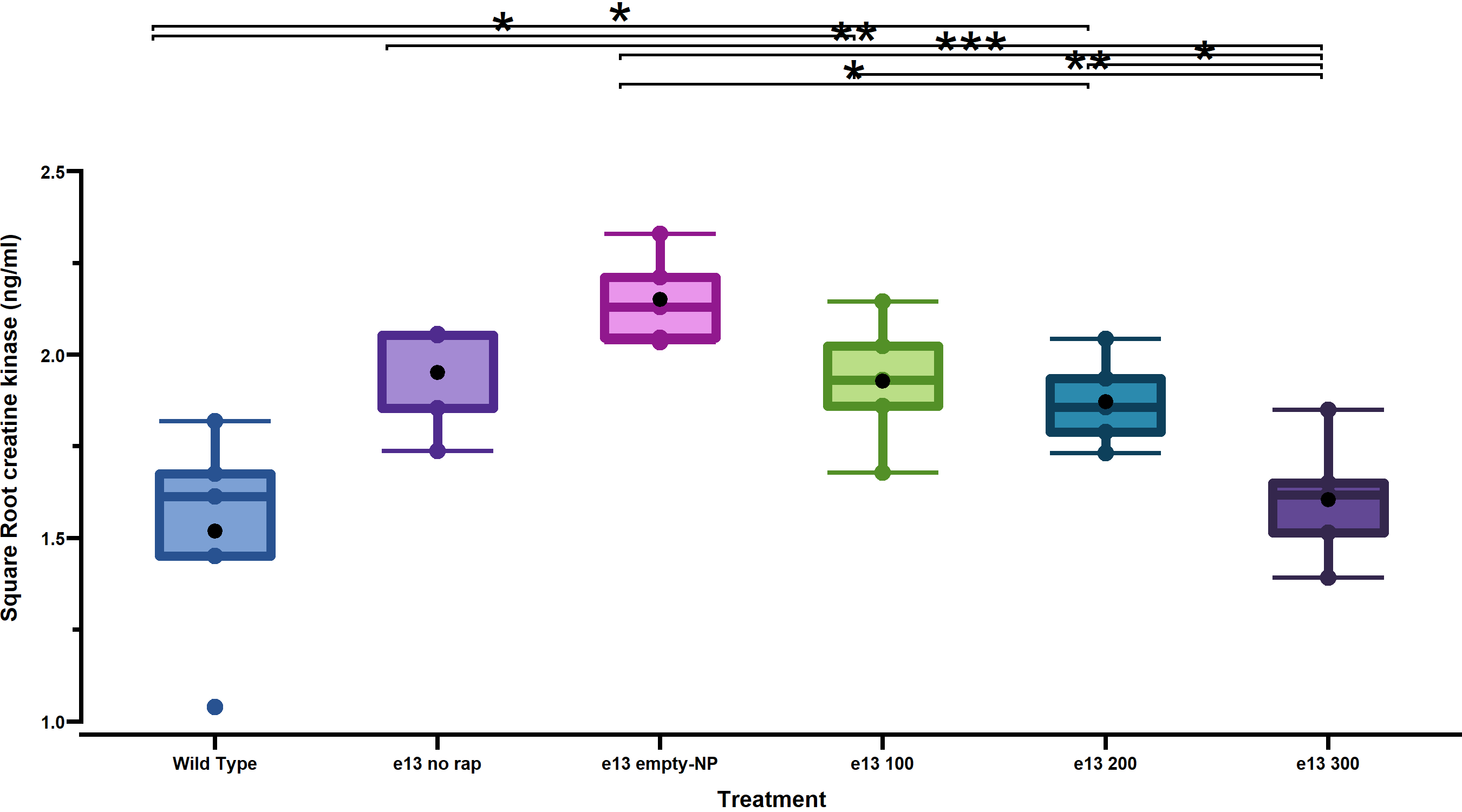


Figure 4.7: Boxplot of Transformed Scale response values at Day 1. The stars above the horizontal bars correspond to the magnitude of the p-value comparing two groups where “\*,” “\*\*,” and “\*\*\*” represent 0.01 <= p-value < 0.05, 0.001 <= p-value < 0.01, and p-value < 0.001, respectively. A p-value summary will only be shown form p-value less than 0.05.

## 4.4 Tables

### 4.4.1 Table 1

|  | | Original Scale | | |
| --- | --- | --- | --- | --- |
| Treatment | Time Points | Mean | Median | SE |
| Wild Type | Overall Average | -0.31 | -0.26 | 0.115 |
| Wild Type | Day 1 | -0.35 | -0.12 | 0.169 |
| e13 no rap | Overall Average | 3.04 | 3.36 | 0.080 |
| e13 no rap | Day 1 | 1.14 | 1.47 | 0.123 |
| e13 empty-NP | Overall Average | 4.10 | 4.16 | 0.092 |
| e13 empty-NP | Day 1 | 1.92 | 1.84 | 0.110 |
| e13 100 | Overall Average | 2.31 | 2.39 | 0.102 |
| e13 100 | Day 1 | 1.02 | 0.99 | 0.128 |
| e13 200 | Overall Average | 1.60 | 1.48 | 0.093 |
| e13 200 | Day 1 | 0.79 | 0.71 | 0.094 |
| e13 300 | Overall Average | 0.71 | 0.63 | 0.085 |
| e13 300 | Day 1 | -0.11 | -0.12 | 0.110 |
| Group level summary on the original scale. | | | | |

### 4.4.2 Table 2

|  | | Transformed Scale | | Back Transformed | | vs. e13 empty-NP | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Treatment | Time Points | Mean | SE | Mean | SE | LSMean Diff (95% CI) | p value |
| Wild Type | Overall Average | 1.52 | 0.086 | 2.32 | 0.26 | -1.07 (-1.299, -0.843) | < 0.001\* |
| Wild Type | Day 1 | 1.52 | 0.116 | 2.31 | 0.35 | -0.63 (-0.933, -0.327) | < 0.001\* |
| e13 no rap | Overall Average | 2.38 | 0.051 | 5.67 | 0.24 | -0.21 (-0.381, -0.047) | 0.012\* |
| e13 no rap | Day 1 | 1.95 | 0.069 | 3.81 | 0.27 | -0.2 (-0.418, 0.02) | 0.079 |
| A Square Root transformation was applied to the data. Difference and CI are estimated using model based LSmean | | | | | | | |

### 4.4.3 Table 3

|  | | Transformed Scale | | Back Transformed | | vs. e13 100 | | vs. e13 200 | | vs. e13 300 | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Treatment | Time Points | Mean | SE | Mean | SE | LSMean Diff (95% CI) | p value | LSMean Diff (95% CI) | p value | LSMean Diff (95% CI) | p value |
| e13 empty-NP | Overall Average | 2.60 | 0.051 | 6.74 | 0.27 | 0.37 (0.173, 0.572) | < 0.001\* | 0.53 (0.327, 0.728) | < 0.001\* | 0.75 (0.552, 0.948) | < 0.001\* |
| e13 empty-NP | Day 1 | 2.15 | 0.069 | 4.62 | 0.30 | 0.22 (-0.035, 0.48) | 0.116 | 0.28 (0.021, 0.536) | 0.029\* | 0.54 (0.287, 0.803) | < 0.001\* |
| e13 100 | Overall Average | 2.22 | 0.051 | 4.94 | 0.23 |  |  | 0.15 (-0.048, 0.358) | 0.176 | 0.38 (0.178, 0.576) | < 0.001\* |
| e13 100 | Day 1 | 1.93 | 0.069 | 3.71 | 0.27 |  |  | 0.06 (-0.202, 0.314) | 0.958 | 0.32 (0.065, 0.579) | 0.008\* |
| e13 200 | Overall Average | 2.07 | 0.051 | 4.28 | 0.21 |  |  |  |  | 0.22 (0.022, 0.423) | 0.026\* |
| e13 200 | Day 1 | 1.87 | 0.069 | 3.50 | 0.26 |  |  |  |  | 0.27 (0.009, 0.523) | 0.04\* |
| e13 300 | Overall Average | 1.85 | 0.051 | 3.41 | 0.19 |  |  |  |  |  |  |
| e13 300 | Day 1 | 1.61 | 0.069 | 2.58 | 0.22 |  |  |  |  |  |  |
| A Square Root transformation was applied to the data. Difference and CI are estimated using model based LSmean | | | | | | | | | | | |

### 4.4.4 Table 4

|  | | vs. Wild Type | | vs. e13 no rap | |
| --- | --- | --- | --- | --- | --- |
| Treatment | Time Points | LSMean Diff (95% CI) | p value | LSMean Diff (95% CI) | p value |
| e13 100 | Overall Average | 0.7 (0.437, 0.961) | < 0.001\* | -0.16 (-0.358, 0.041) | 0.158 |
| e13 100 | Day 1 | 0.41 (0.06, 0.755) | 0.016\* | -0.02 (-0.281, 0.234) | 1 |
| e13 200 | Overall Average | 0.54 (0.281, 0.807) | < 0.001\* | -0.31 (-0.515, -0.112) | < 0.001\* |
| e13 200 | Day 1 | 0.35 (0.004, 0.699) | 0.046\* | -0.08 (-0.337, 0.178) | 0.91 |
| e13 300 | Overall Average | 0.32 (0.059, 0.584) | 0.012\* | -0.54 (-0.734, -0.337) | < 0.001\* |
| e13 300 | Day 1 | 0.09 (-0.262, 0.433) | 0.943 | -0.35 (-0.604, -0.088) | 0.004\* |
| A Square Root transformation was applied to the data. Difference and CI are estimated using model based LSmean | | | | | |

# 5 References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, and Richard Iannone. 2020. *Rmarkdown: Dynamic Documents for r*. <https://github.com/rstudio/rmarkdown>.

Dawson, Charlotte. 2021. *Ggprism: A ’Ggplot2’ Extension Inspired by ’GraphPad Prism’*. <https://doi.org/10.5281/zenodo.4556067>.

Edwards, Don, and Jack J. Berry. 1987. “The Efficiency of Simulation-Based Multiple Comparisons.” *Biometrics* 43 (4): 913–28. <http://www.jstor.org/stable/2531545>.

Lenth, Russell V. 2022. *Emmeans: Estimated Marginal Means, Aka Least-Squares Means*. <https://CRAN.R-project.org/package=emmeans>.

Liu, Wei, Mortaza Jamshidian, Ying Zhang, Frank Bretz, and Xiaoliang Han. 2007. “Some New Methods for the Comparison of Two Linear Regression Models.” *Journal of Statistical Planning and Inference* 137 (1): 57–67. https://doi.org/<https://doi.org/10.1016/j.jspi.2005.09.007>.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. 2013. *Nlme: Linear and Nonlinear Mixed Effects Models*.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

RStudio, Inc. 2013. *Easy Web Applications in r.*

Sievert, Carson. 2020. *Interactive Web-Based Data Visualization with r, Plotly, and Shiny*. Chapman; Hall/CRC. <https://plotly-r.com>.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>.

# 6 Appendix: Workflow Illustration

