**Methods for Data Analysis – Project**

by

**Sundar Srinivasan**

**Introduction**

The data analyzed for this project was collected as part of a series of experiments performed at the University of Washington (please see: ProjectData\_SSF.csv).

Unfortunately, as the data is unpublished, the subject IDs, treatments, independent and dependent variables had to be given anonymous names prior to posting on github.

Per the dogma of our field, the outcomes (labeled Outcome\_1, Outcome\_2, Outcome\_3) are typically expected to be explained/predicted by the independent variable Indi\_1. However, the dogma does not explain more recent data, and interventions based upon the dogma have broadly proved ineffective. More recently, we have observed in pilot trials that under different experimental treatments, outcome variables are rapidly altered despite only a mild defect in the ‘dogma’ variable Indi\_1. Interestingly, subjects undergoing treatments displayed more pronounced alteration in their gait. To further probe these preliminary observations and to re-examine the dogma, a broader experiment was performed to test the hypothesis that:

**Alterations in gait, not in Indi\_1, associated with the treatments predicts outcome variables**.

To test the hypothesis, groups of subjects (n=8/grp) were exposed to 5 different experimental conditions: Treatment1 – Treatment5. Following treatment, the outcome measures (Outcome\_1, Outcome\_2, and Outcome\_3) were measured at day 0, 5, 12, 28. Additionally, the ‘predictive’ independent variates: Indi\_1 (dogma) and Indi\_2 were also determined on days 1, 5, 12, 28. Lastly, all subjects were accessed (on day 0, 5, 12, 28) over trials (n = 10-20/subject/day) for kinematic traits where video images of gait was recorded (~60 Hz, ~ 5 mins/trial/subject) and 5 measures of gait were extracted from the images: area (of subject), center of mass in x (CMx), center of mass in y (CMy), height and length of the subject. These measures were then converted from the time domain to the frequency domain via fast Fourier transform (fft). These data (i.e., Indi\_1, gait fft components, Indi\_2) were then used to test our hypothesis primarily using exploratory and predictive linear regressions.

**Methods:**

All data analysis was carried out in the R programming language (Main\_ProjectSS.R). As a first step towards analyzing the data, a qualitative (plots of mean, s.d.) exploration of Indi\_1, Indi\_2, and fft components of gait, and outcome variables were performed over time and across treatment type.

Secondly, given that measures of gait extracted from video images were likely dependent, an attempt was made to determine which measure of gait (i.e., area, CMx, CMy, length, height) could ‘best’ categorize the treatments by type. For this, multinomial logistic regression was used to classify experimental conditions using fft components derived from the 5 different measures of gait (i.e., area, CMx, CMy, length, height). Prior to classification via regression, data transformations were performed where measurements obtained on the different experimental days (i.e., day 0, 5, 12, 28) were treated as independent variates. This was achieved via a join operation on mouse id and experimental treatment. However, as the number of trials performed in assessing gait varied between subjects and day of trial (from n = 10 -20), a representative trial was defined as one with the minimum Euclidian distance to the mean of the fft components across trials. For the classification, multinomial lasso regression and cross validation was implemented and the measure of gait selected for further analysis was one that most accurately classified treatment types. This classification was also performed with a view to reducing the computational cost associated with the exploratory and predictive regressions that were subsequently performed to test the broad hypothesis (~ 5 X reduction in wall-clock times).

Third, exploratory linear regressions were performed to examine whether Indi\_1 (dogma trait), Indi\_2, or gait components could explain the variation in outcome variables observed experimentally. To follow-up the initial results, regression models were built to include treatment type and day of experimental observation as additional independent variates. Lastly, models were also constructed to explore interactions between the primary observations (i.e., Indi\_1, Indi\_2, gait components) with treatment type and the day when experimental measurements were obtained. In analyses with more than one independent variate, model selection was performed via AIC using the step-wise algorithm (in R).

Lastly, linear regressions were performed to examine whether Indi\_1, Indi\_2 and/or gait components without and with day of experiment as independent factor could predict outcome measures. This analysis was performed to further test our hypothesis. Additionally, by not including treatment type as an independent factor, the model’s ability to predict outcomes induced by yet to be implemented experimental conditions was also indirectly evaluated. For the analysis, the dataset was separated into a training and test subset using a 70-30 random split. In the cases where models were constructed with more than one independent variate, the ‘optimal’ form of the model, given the training data, was obtained using lasso-regression and cross-validation. Post-fitting of the model to the training data, it was then used to predict outcome measures in the test data subset.

**Results**

Qualitative exploration of Indi\_1, Indi\_2, gait components, and outcome measure changes over time and treatment type (folder: “Figures”) yielded an initial understanding of variations in independent and dependent measures (Figure 1). For instance, outcome measures were altered over time and varied with treatment type. While Indi\_2 alterations appeared to undergo similar trends over time and treatment type, the fft components of gait (e.g., 1 Hz; Figure 1) and Indi\_1 variations were less reflective of alterations in outcome measures.

The gait measure that most accurately classified experimental (treatment) type was center of mass in the y-direction (CMy). However, due to the stochasticity inherent in the lasso regression and cross-validation, other measures (e.g., length) were also, albeit less frequently, identified as the measure that most accurately classified experimental conditions. As such, subsequent exploratory and predictive regression partitioned the data by utilizing the measure identified to provide the most accurate classification for that particular iteration of the analysis.

Exploratory linear regressions suggested that both Indi\_1 and fft components (individually or collectively) of gait poorly explained the variance in outcome measures (< 10%); Table 1; Figure 2). In contrast, Indi\_2 (in and of itself) explained 43% and 38% of the variance in Outcome\_1 and Outcome\_3, and 12% of the variance in Outcome\_2. Models selected from the full spectrum of independent variables, i.e., Indi\_1, all gait components, Indi\_2, and their interactions between treatment type and day, resulted in the highest explanation of variance in Outcome\_1 (81%), Outcome\_3 (72%), and Outcome\_2 (50%; Table 1).

Lastly, when linear regressions were used to predict experimental outcomes, it was found that Indi\_1 poorly (but significantly, p < 0.05) explained the variance (≤ 12%, Table 2; Figure 3). However, gait components failed to explain the variance in outcome measures (p > 0.05). In contrast, Indi\_2 could predict 44% of the variance in Outcome\_1, 39% of the variance in Outcome\_3 and 11% of the variance in Outcome\_2. Lastly, models that included the primary measurements of Indi\_1, Indi\_2 and gait components, as well as interactions with day of measurement predicted 65% of the variance in Outcome\_1, 44% in Outcome\_3 and 20% in Outcome\_2 (Table 2).

**Discussion**

The analysis suggests that alterations in gait or in Indi\_1 induced by treatments were poor predictors of outcome measures. While unexpected, given the dogma as well as the distinctive qualitative changes in gait observed in subject undergoing these treatments, our analysis served to reject our broad hypothesis.

Surprisingly, Indi\_2 was identified as the most significant explanatory and predictive variate, particularly in the context of experimental outcomes observed in anatomically similar locations (i.e., 44% and 39% for Outcome\_1, Outcome\_3, respectively). Further including gait components, Indi\_1 and day of measurement as independent variates improved prediction of Outcome\_1, 3 by an additional 20% and 5%, respectively. Extension of this analysis could permit an understanding of the contributory influence of each independent variate when superimposed upon the predictions enabled by Indi\_2.

The identification of Indi\_2 as a primary predictor of Outcome\_1, 3 provides opportunities for further study. This is especially relevant in the context that the experimental treatments altered Indi\_2 function only for treatment # 3 directly adjacent to the anatomical location where outcomes were measured. This suggests some form of higher level biological integration of perturbations caused by treatments, and indicates the need to acquire and include more nuanced and/or refined experimental information on Indi\_2.

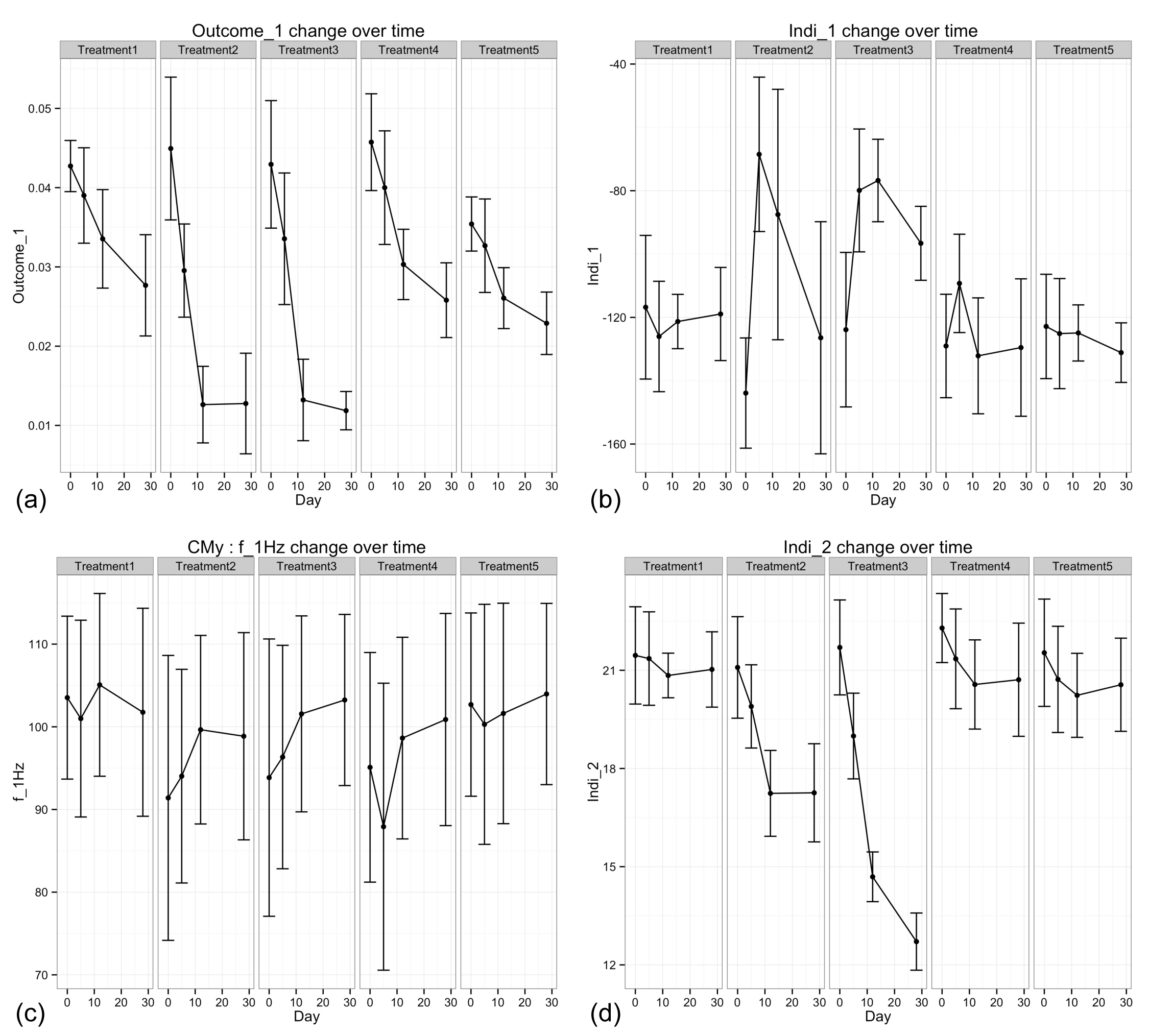
In conclusion, this analysis rejected our broad hypothesis that alterations in gait associated with the treatments predicts outcome measures. Importantly, our analysis rejects the dogma of our field, and suggests that Indi\_1 cannot explain the outcome measurements. For the first time, this analysis also suggested that Indi\_2 was the primary predictor of outcome measures. Inclusion of measures of gait, Indi\_1 and time of measurement alongside Indi\_2 does, however, improve ability to predict outcome measures.

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | Outcome\_1 | Outcome\_2 | Outcome\_3 |
| Indi\_1 | 0.069 | 0.0 | 0.085 |
| Indi\_1: TreatmentType | 0.171 | 0.077 | 0.325 |
| Indi\_1: Day | 0.526 | 0.209 | 0.284 |
| Indi\_1: TreatmentType: Day | 0.665 | 0.322 | 0.599 |
| f\_2Hz | 0.012 | 0.001 | 0.002 |
| f\_2Hz: TreatmentType | 0.203 | 0.11 | 0.38 |
| f\_2Hz: Day | 0.438 | 0.131 | 0.162 |
| f\_2Hz: TreatmentType: Day | 0.651 | 0.303 | 0.59 |
| Indi\_2 | 0.427 | 0.118 | 0.375 |
| Indi\_2:TreatmentType | 0.525 | 0.289 | 0.561 |
| Indi\_2:Day | 0.618 | 0.205 | 0.416 |
| Indi\_2:TreatmentType:Day | 0.734 | 0.334 | 0.628 |
| f\_1Hz to f\_30 Hz | 0.019 | 0.013 | 0.013 |
| f\_1Hz to f\_30 Hz: TreatmentType | 0.237 | 0.134 | 0.399 |
| f\_1Hz to f\_30 Hz: Day | 0.443 | 0.15 | 0.181 |
| f\_1Hz to f\_30 Hz: TreatmentType: Day | 0.694 | 0.338 | 0.625 |
| **f\_1Hz to f\_30 Hz: Indi\_1: Indi\_2: TreatmentType: Day** | **0.81** | **0.501** | **0.724** |

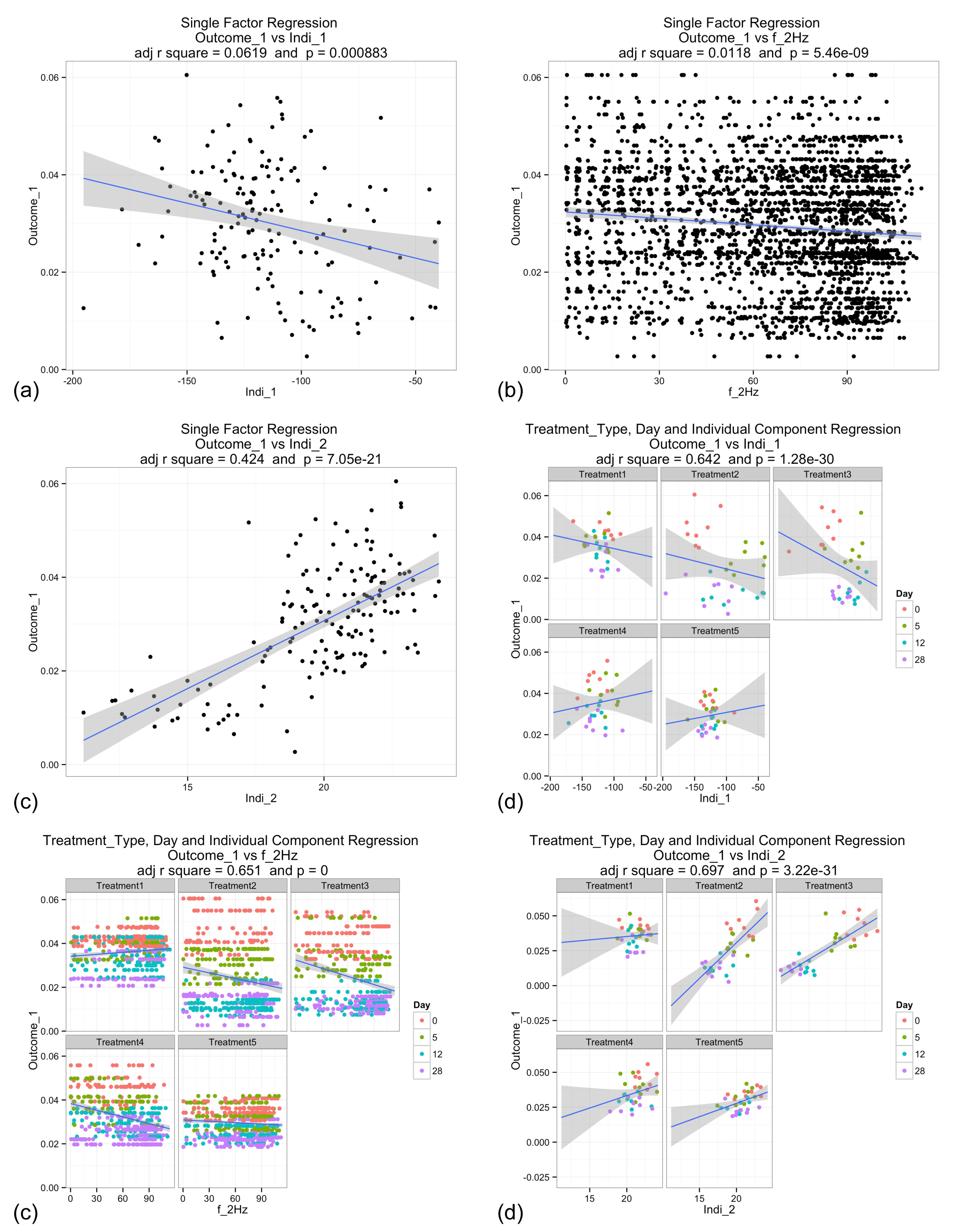
**Table 1:** r2 values for explanatory regression models by dependent variable (i.e., Outcome\_1 to \_3)

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Outcome\_1 | Outcome\_2 | Outcome\_3 |
| Indi\_1 | 0.09 ± 0.06 | 0.03 ± 0.03# | 0.12 ± 0.09 |
| Indi\_2 | 0.44 ± 0.08 | 0.11 ± 0.05 | 0.39 ± 0.06 |
| Indi\_1:Day | 0.52 ± 0.05 | 0.16 ± 0.09 | 0.31 ± 0.11 |
| Indi\_2:Day | 0.60 ± 0.06 | 0.14 ± 0.08 | 0.40 ± 0.06 |
| Indi\_1:Indi\_2:Day | 0.60 ± 0.06 | 0.15 ± 0.06 | 0.39 ± 0.07 |
| Gait (1-30 Hz) | 0.005 ± 0.002# | 0.002 ± 0.003# | 0.002 ± 0.003# |
| Gait (1-30 Hz): Day | 0.44 ± 0.02 | 0.12 ± 0.01 | 0.15 ± 0.02 |
| Gait (1-30 Hz): Indi\_1 : Day | 0.52 ± 0.01 | 0.17 ± 0.02 | 0.26 ± 0.02 |
| Gait (1-30 Hz): Indi\_2: Day | 0.65 ± 0.01 | 0.19 ± 0.02 | 0.42 ± 0.01 |
| Gait (1-30 Hz): Indi\_1: Indi\_2: Day | 0.65 ± 0.01 | 0.20 ± 0.02 | 0.44 ± 0.02 |

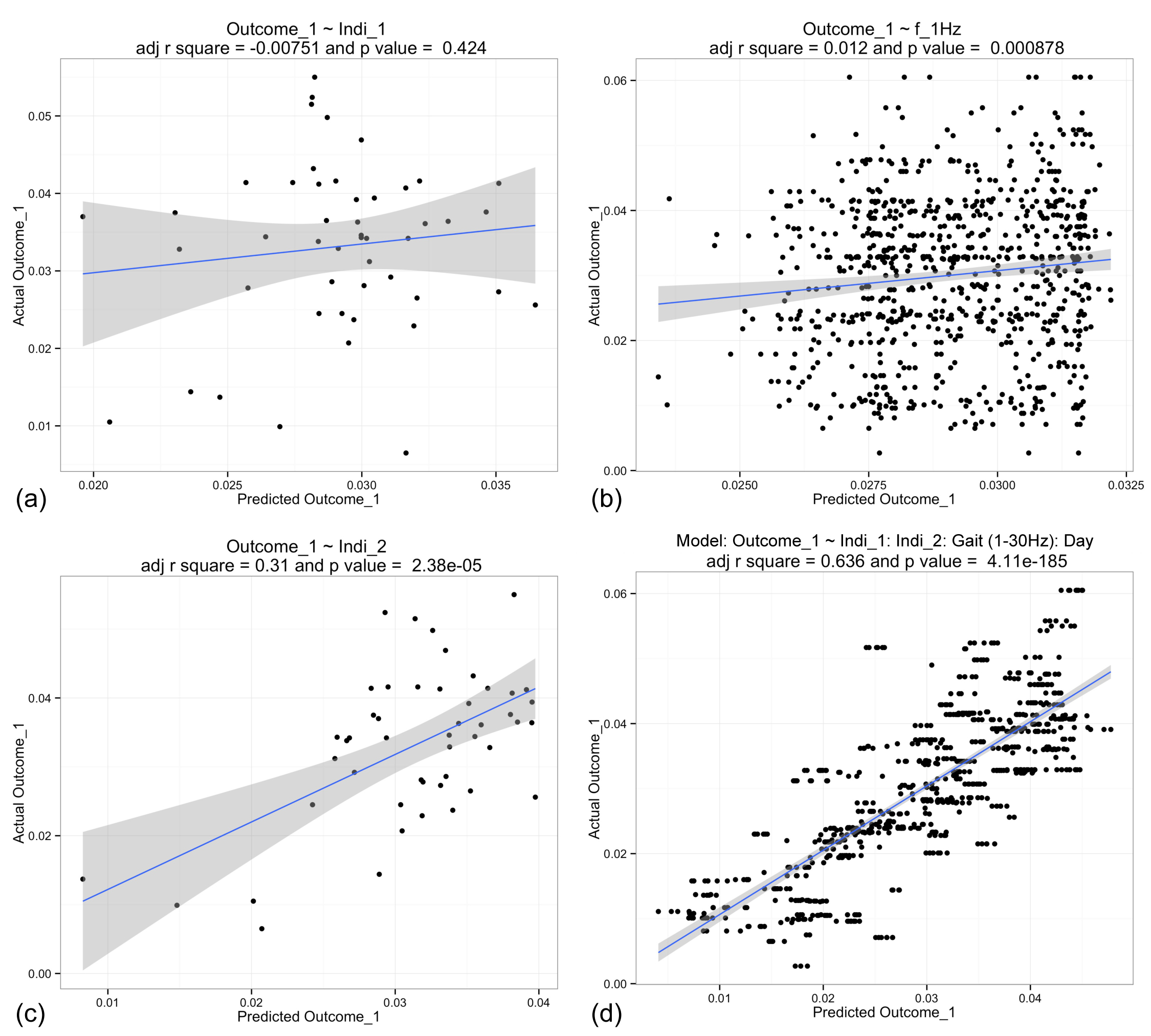
**Table 2:** r2 values (mean ± s.d.) for predicted vs measured Outcomes (# indicates r2 values that do not attain significance, i.e., p > 0.05).



***Figure 1:*** Illustration of Outcome\_1 (a), Indi\_1 (b), 1-Hz component of gait derived from CMy (c), and Indi\_2 across time and treatment type (d; mean +/- s.d.).



***Figure 2:*** Linear regression of Outcome\_1 vs Indi\_1 (a), the 2-Hz fft component of gait (b), Indi\_2 (c), interactions between treatment type, day and Indi\_1 (d), interactions between treatment type, day and the 2-Hz component (e), and interactions between treatment type, day and Indi\_2 (f).



***Figure 3:*** Linear regression of Outcome\_1 from the test data set vs Indi\_1 (a), vs gait components (i.e., 1 – 30 Hz; b), Indi\_2 (c), and Indi\_1, gait components, Indi\_2 and day (d).