Examining BotB treatment effects over time for CD patients

Lily Bai Nathalie Blasco Yihao Peng Lauren Rackley Sirui Wu

2025-11-27

Introduction

The purpose of this project is to examine the effects of botulism toxin type B (BotB) to treat cervical dystonia over time. Cervical dystonia (CD) is a chronic neurological disorder, in which patients have painful involuntary contractions in neck muscles. CD is more prevalent in women (Jankovic et al., 2023). The prevalence of CD is estimated to be 28-183 cases per million. The data comes from a multicenter randomized clinical trial for cervical dystonia patients with 9 U.S. sites. Botulism toxin types A and B are first-line treatments for CD (Wetmore et al., 2025). The treatment groups included in the study were placebo, 5000 U BotB, and 10000 U BotB. The response variable is Toronto Western Spasmodic Torticollis Rating Scale (TWSTRS) total score, which ranges from 0 to 87 and is comprised of disability (0-32), pain (0-20), and severity (0-35) subscores. Only total score is included in this data. The TWSTRS score was measured at week 0 (baseline), and 2,4,8,12, and 16 weeks after treatment start. Site is included in the dataset but no further details about site were included in the available dataset documentation.

Methods

Study Population

To be included in the study, patients must have had CD for at least 1 year involving two or more muscles and CD continued response to BotA treatment. Patients also had to have TWSTRS score $\geq$ 20 with severity score $\geq$ 10, disability score $\geq$ 3, and pain score $\geq$ 1. Patients were at least 18 years old, minimum weight of 46 kg, and had acceptable phsyical and neurological exams as well as labs. Patients who had neurotoxin injection in the last 4 months were exlcuded and could not have participated in previous BotB trials (Brashear et al., 1999). The study included 109 patients (67 (61%)) females. The mean age was 55.5 (12.1) years. Median age was 56.0 years. The mean TWSTRS score at baseline was 45.7 (9.7). Demographic and baseline characteristics were similar among the three treatment groups (Table 1). It was not known if the patients received prior BotB treatments. Information about the randomization schedule was not provided.

Statistical Analyses

Number of observations, mean, median, standard deviation (SD), minimum (min) and maximum (max) were provided for age. Mean and SD were calculated for TWSTRS score at baseline. Frequencies and percentages were reported for categorical variables. GLM, GLMM, and GEE models were fit using TWSTRS total score as the response variable. For GLM, age, sex, week, and treatment were used as covariates with male and placebo being the reference groups. For the GEE and GLMM models, week, age, sex, treatment, and treatment $\times$ week interaction were included in the model again with male and placebo being the reference groups. Both random slope and random intercept models were fit. ANOVA was used to compare the random slope and random intercept models. A Gaussian link function was used for all three model types. Statistical analyses were performed using R (version 4.4.2, R Core Team, 2024).

Results

Table 1: Summary of Demographic and Baseline Characteristics

Characteristic	Overall N = 109	Placebo N = 36	BotB 5000 U N = 36	BotB 10000 U N = 37	p-value
Sex, n(%)					0.0706
Male	42 (39%)	15 (42%)	18 (50%)	9 (24%)	
Female	67 (61%)	21 (58%)	18 (50%)	28 (76%)	
Age (years)					0.6198
N	109.0	36.0	36.0	37.0	
Mean (SD)	55.5 (12.1)	53.8 (12.3)	57.1 (12.4)	55.7 (11.8)	
Median	56.0	55.5	57.0	54.0	
Min, Max	26.0, 83.0	26.0, 79.0	35.0, 83.0	34.0, 76.0	
TWSTRS total score at baseline					0.3307
Mean (SD)	45.7 (9.7)	43.6 (9.0)	46.4 (10.4)	46.9 (9.6)	

[1] BotB = botulinum toxin type B; TWSTRS = Toronto Western Spasmodic Torticollis Rating Scale.

[2] Pearson's Chi-squared test; Kruskal-Wallis rank sum test

Generalized Linear Model (GLM)

A generalized linear model (GLM model) using a Gaussian link was fit including week, treatment, age, and sex as predictors (Table 3).

Table 2: GLM Model Summary (Gaussian Link)

term	estimate	std.error	statistic	p.value
(Intercept)	37.673	2.582	14.588	0.000
week	0.235	0.089	2.631	0.009
treat5000 U	0.005	1.248	0.004	0.997
treat10000 U	-0.347	1.251	-0.277	0.782
age	0.017	0.042	0.396	0.692
sexFemale	2.183	1.071	2.038	0.042

The coefficient for week was statistically significant, indicating an overall linear trend in TWSTRS score over time. the coefficient for sex was also statistically significant, indicating a difference in TWSTRS score between females and males. Male is used as the reference group, so females have a TWSTRS score that is 2.18 points higher than males, on average holding everything else constant. However, because the GLM assumes independence of observations, the repeated measures within individuals violate this assumption. As a result, the standard errors and p-values may be underestimated, and inference should be interpreted with caution. This motivates the subsequent use of correlation-aware models such as GEE and GLMM.

GLM Model Diagnostics

Diagnostics were assessed for the GLM model.

Figure 1: Residuals vs Fitted Plot

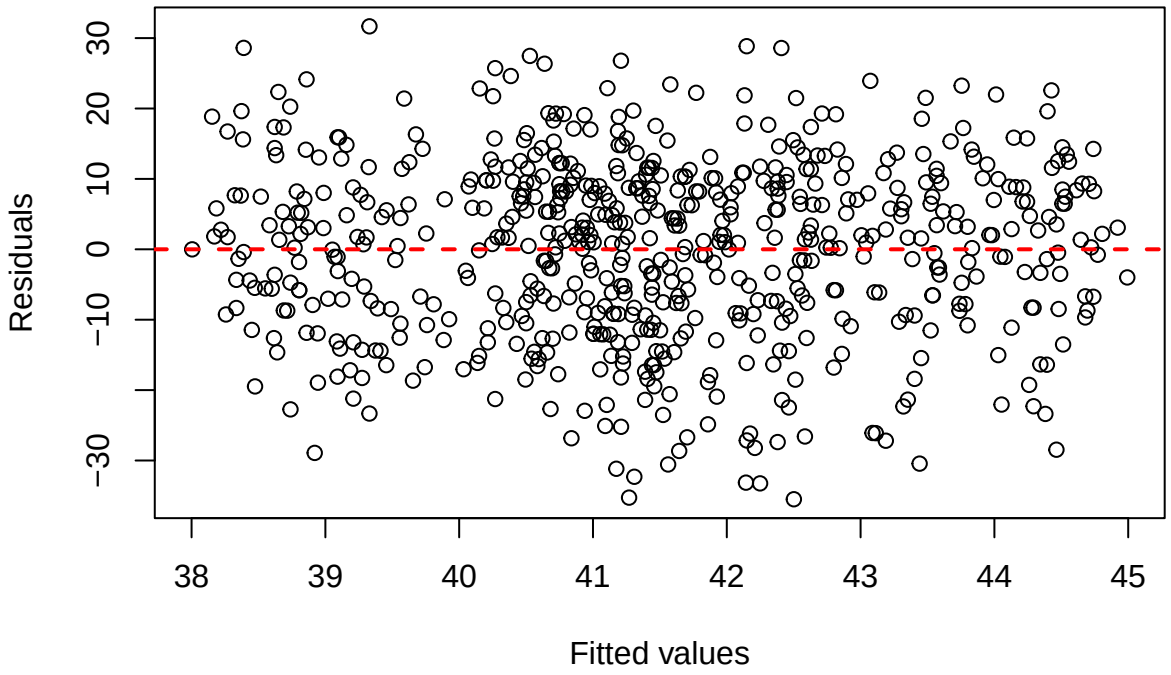


Figure 1 shows the residuals versus the fitted values. The absence of strong patterns or fanning suggests that the linearity and homoscedasticity assumptions are reasonably met.

Figure 2: Q−Q Plot

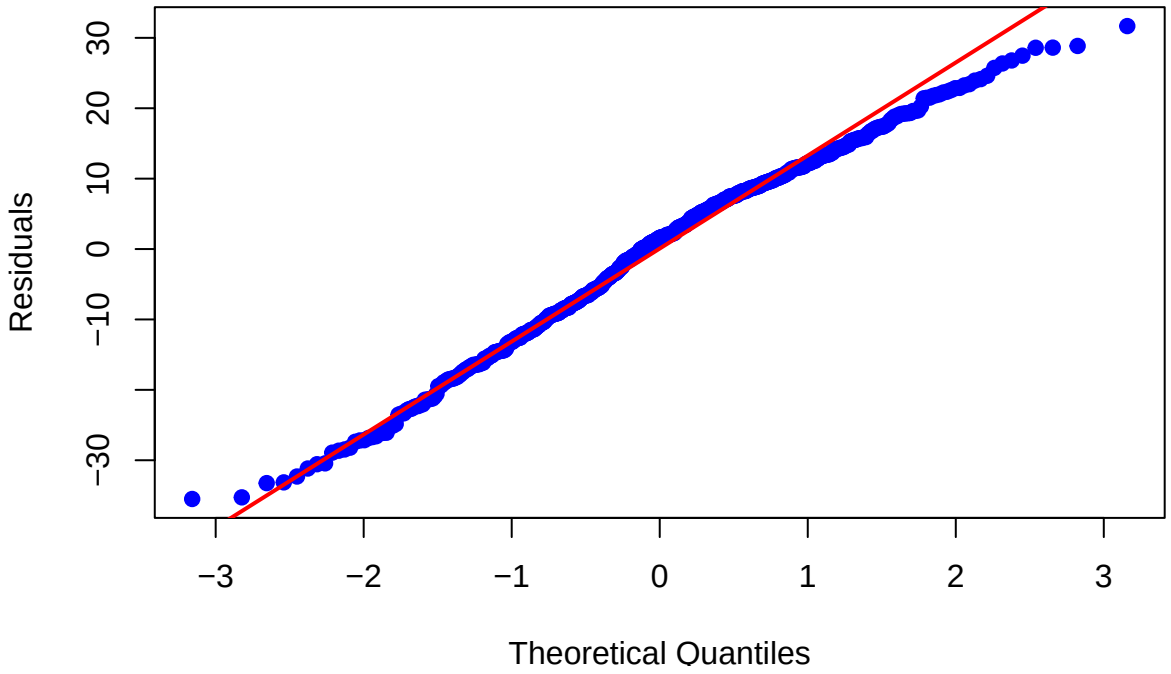
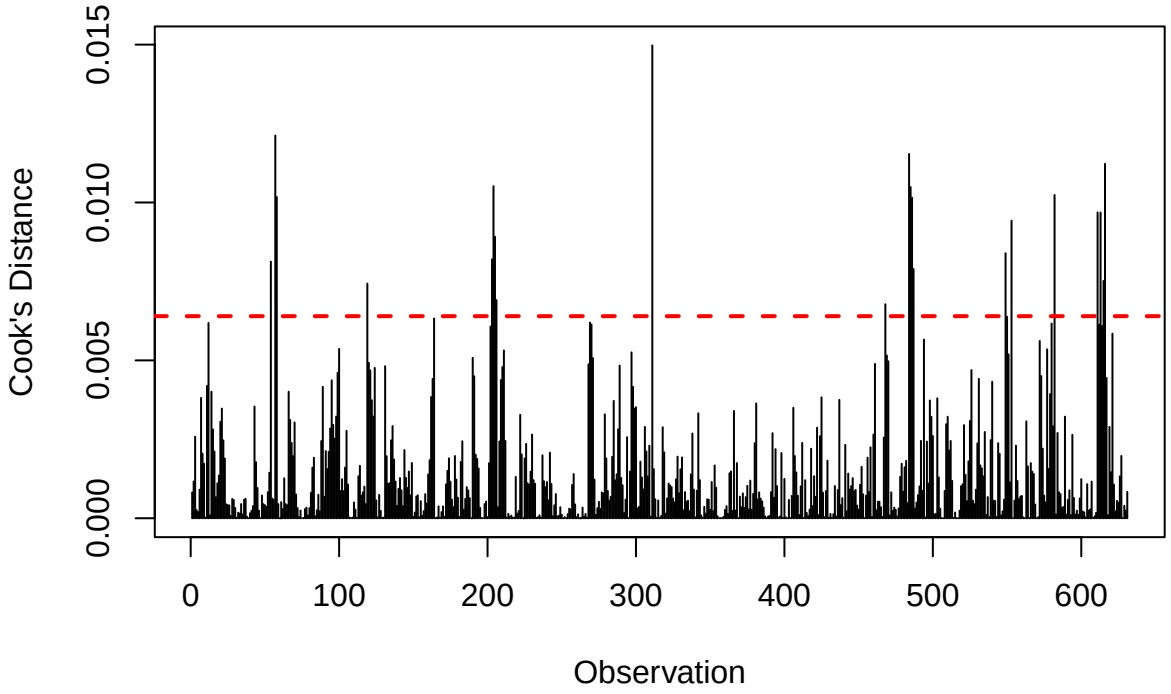


Figure 2 presents a QQ plot of the residuals, which largely follow the 45 degree reference line with mild deviations in the tails. This indicates that the normality assumption is approximately satisfied.

Figure 3: Cook's Distance Plot



Finally, figure 3 displays Cook's distance for all observations. The conventional threshold, 4/(n-k-1), where n is the number of observations and k is the number of predictors, was used to assess the influence of the observations. Several observations exceeded the threshold and were considered potentially influential and warranted further investigation. Therefore, a second GLM was fit, excluding those observations.

Table 3: GLM Model Summary (Excluding Influential Observations)

term	estimate	std.error	statistic	p.value
(Intercept)	36.597	2.449	14.946	0.000
week	0.276	0.084	3.284	0.001
treat5000 U	0.178	1.182	0.151	0.880
treat10000 U	-1.159	1.160	-1.000	0.318
age	0.031	0.040	0.784	0.433
sexFemale	3.217	1.008	3.193	0.001

Excluding influential observations resulted in modest shifts in the parameter estimates and slightly improved precision. Notably, the effect of week and sex remained statistically significant. Overall, excluding influential points slightly adjusted the estimates and improved precision, but the main patterns of association were consistent with the original GLM. As with the previous model, the GLM assumptions are not appropriate for longitudinal data with correlated repeated measures. The following GEE and GLMM analyses address this limitation by explicitly modeling within-subject correlation and random effects.

Generalized Estimating Equation (GEE)

A GEE model with Gaussian family, identity link, and an exchangeable correlation structure. The mean model included week, treatment, treatment-by-week interaction, and baseline age and sex, with Placebo as the reference treatment group.

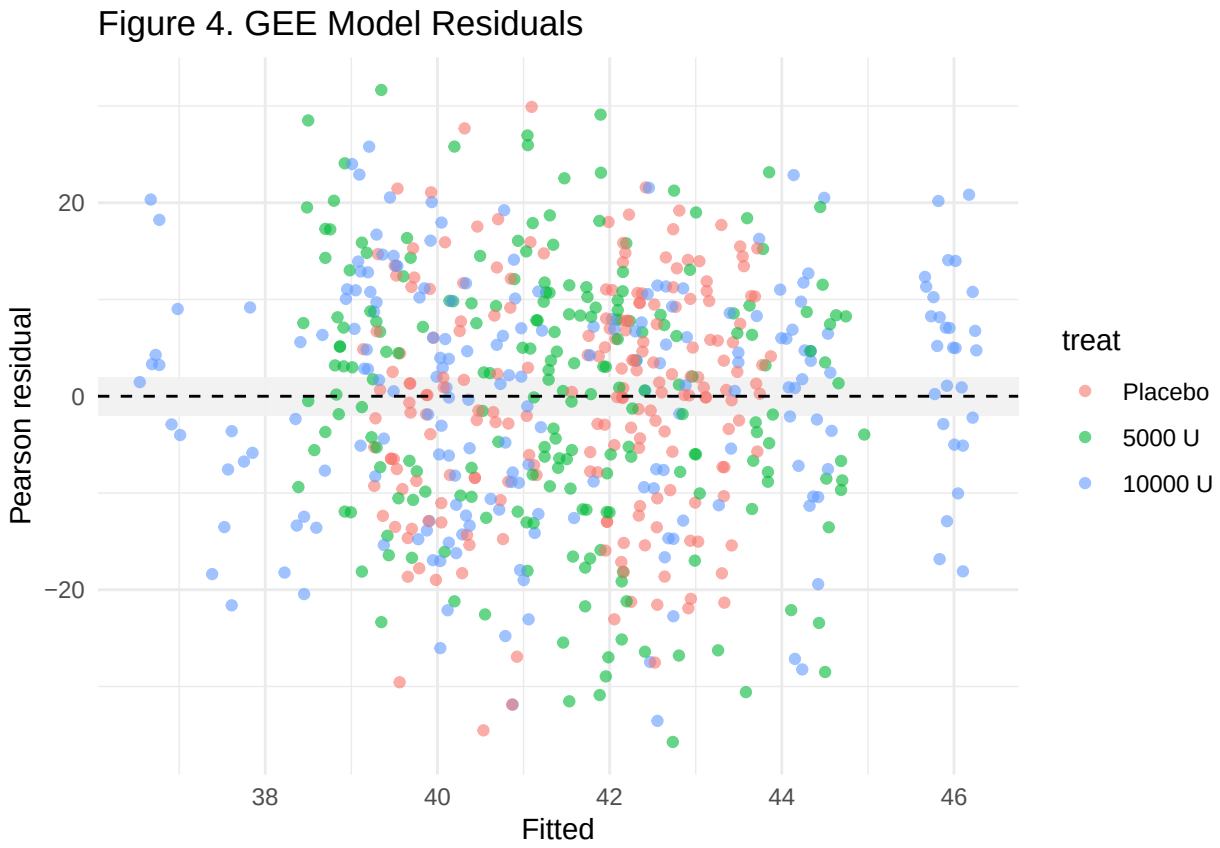


Figure 4 shows Pearson residuals plotted against the fitted TWSTRS values with points colored by treatment group. The residuals are centered around zero and do not show a strong funnel shape, supporting the constant-variance assumption. However, many residuals fall outside ± 2, indicating wide individual variation in TWSTRS trajectories and suggesting that the simple Gaussian working model does not fully capture the variance structure. Even so, because the GEE analysis used robust standard errors, the uncertainty in the regression coefficients is driven by the observed variability of the residuals within subjects, rather than depending on the working variance and correlation.

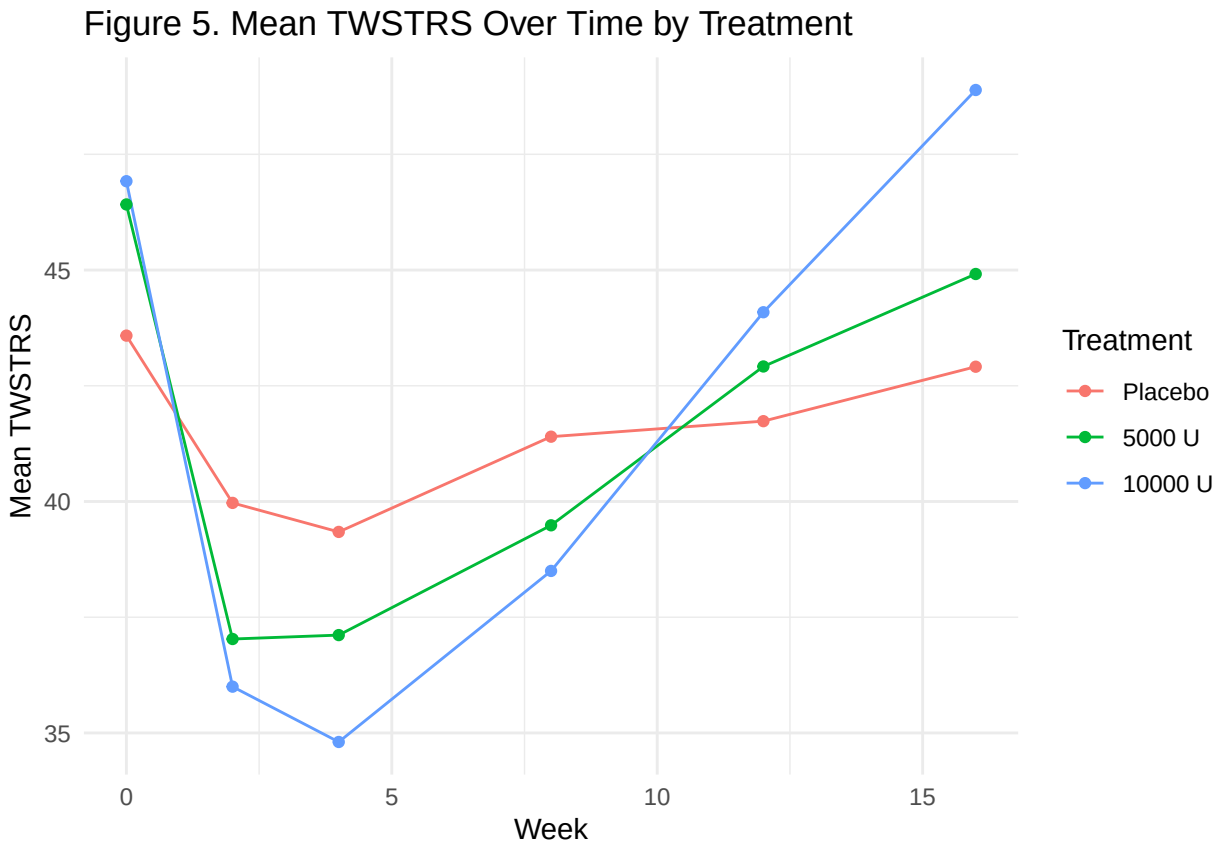


Figure 5 shows mean TWSTRS over time by treatment group. All groups show decreases from baseline through weeks 2–4, suggesting early improvement, followed by increases after week 4. Around week 10, both active treatment groups begin to show higher TWSTRS scores than the placebo group.

Table 4: GEE Model Summary

term	estimate	std.error	statistic	p.value
(Intercept)	38.698	4.760	66.095	0.000
week	0.097	0.114	0.731	0.392
treat5000 U	-0.810	2.758	0.086	0.769
treat10000 U	-2.738	2.372	1.333	0.248
age	0.014	0.078	0.033	0.855
sexFemale	2.494	2.117	1.388	0.239
week:treat5000 U	0.115	0.133	0.745	0.388
week:treat10000 U	0.323	0.146	4.899	0.027

Table 4 summarizes the GEE estimates for TWSTRS over time by treatment, adjusted for age and sex. The main effects for treat5000U (-0.810, $p = 0.769$) and treat10000U (-2.738, $p = 0.248$) compare baseline TWSTRS at week 0 with placebo. Both estimates are negative but not statistically significant, indicating no clear baseline differences in mean TWSTRS between treatment groups and placebo. The treatment-by-week interaction terms describe how each treatment's trajectory differs from placebo over time. For 5000U, the additional slope relative to placebo is small and not significant (0.115, $p = 0.388$), suggesting that its average linear trend over time is similar to placebo. For 10000U, the additional slope is larger and statistically significant (0.323, $p = 0.027$), indicating an average increase in TWSTRS over 16 weeks compared with placebo. Because higher TWSTRS scores reflect worse symptoms, these results imply that low-dose regimen behaves similarly to placebo, while the high-dose regimen is associated with faster worsening over time.

Overall, the GEE model provides a reasonable approach for assessing whether treatment affects population-average TWSTRS trajectories over time, while recognizing that the true time course is somewhat non-linear and that estimated treatment effects are relatively small.

Generalized Linear Mixed Effects Model (GLMM)

While the GEE model provides population-averaged estimates, we implemented a Generalized Linear Mixed Model (GLMM) to model subject-specific heterogeneity. By including a random intercept, the GLMM partitions the total variance into between-subject and within-subject components, allowing us to estimate treatment trajectories conditional on an individual patient's unique baseline severity. Additionally, because GLMM uses likelihood-based estimation, they provide valid inference under the Missing at Random (MAR) assumption.

Table 5: GLMM Model Summary

effect	term	estimate	std.error	statistic	df	p.value	conf.low	conf.high
fixed	(Intercept)	38.698	5.248	7.373	107.898	0.000	28.295	49.101
fixed	treat5000 U	-0.810	2.748	-0.295	128.543	0.769	-6.246	4.627
fixed	treat10000 U	-2.738	2.738	-1.000	128.049	0.319	-8.155	2.679
fixed	week	0.097	0.089	1.089	522.247	0.277	-0.078	0.273
fixed	age	0.014	0.088	0.162	104.208	0.871	-0.159	0.188
fixed	sexFemale	2.495	2.216	1.126	104.884	0.263	-1.898	6.888
fixed	treat5000 U:week	0.115	0.125	0.916	521.089	0.360	-0.132	0.361
fixed	treat10000 U:week	0.323	0.125	2.582	521.609	0.010	0.077	0.569

Table 5 presents the fixed effect estimates from the linear mixed model. Consistent with the previous GEE analysis, the point estimates for the fixed effects are identical, as expected for a linear model with a Gaussian distribution. However, the standard errors and p-values differ slightly reflecting the distinction between model-based (GLMM) and robust (GEE) variance estimation.

At baseline, there were no statistically significant differences in disease severity between the treatment arms and the placebo group. Neither age nor sex are significant predictors for TWSTRS. The primary finding is the significant interaction between the 10000 U treatment and week ($\beta = 0.323, SE = 0.125, p = 0.010$). This confirms that, conditional on the individual patient, the high-dose group experienced a significantly faster rate of worsening, increasing by approximately 0.32 points per week relative to the placebo group.

npar	**AIC**	**BIC**	**logLik**	**-2*log(L)**	**Chisq**	**Df**	**Pr(>Chisq)**
10	4.57e+03	4.62e+03	-2.28e+03	4.55e+03			
12	4.58e+03	4.63e+03	-2.28e+03	4.55e+03	1.51	2	0.471

The ANOVA analysis shows that random slope model is not better than the random intercept model.

Figure 6. Conditional Predictions of TWSTRS Score at Different Times

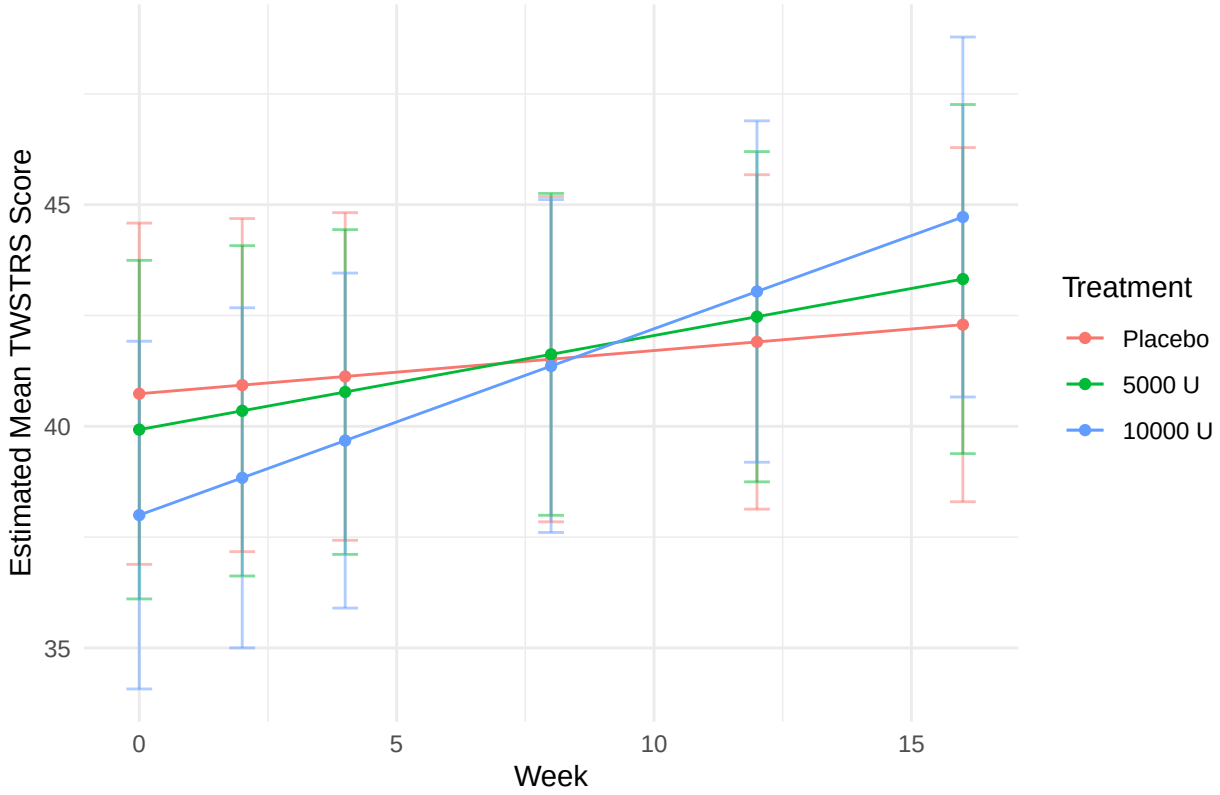


Figure 6 illustrates the model-predicted trajectories for TWSTRS across the three treatment groups. The blue line representing 10000 U treatment and the red line for placebo cross each other, showing the significant interaction term between 10000 U and week. The relationship between treatment and score depends on time.

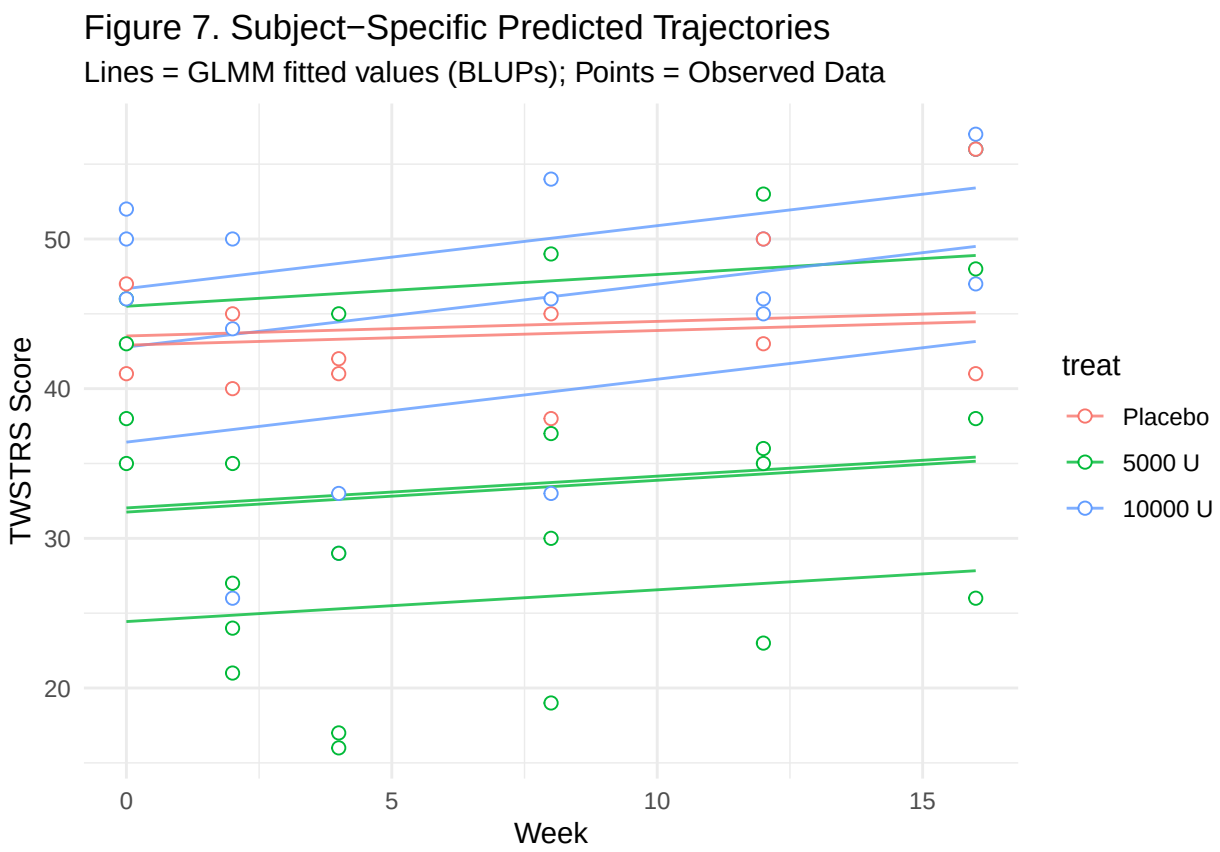


Figure 7 displays the observed TWSTRS scores (points) overlaying the GLMM fitted trajectories (lines) for a random sample of 12 subjects. The heterogeneity at baseline confirms the necessity of including a random intercept for each subject to account for patient-specific heterogeneity, suggesting that much of the variance arises from pre-existing differences between patients.

The relationship between the solid lines and scattered points demonstrates how the model generates Best Linear Unbiased Predictions (BLUPs) for each subject. Instead of overfitting to the noisy week-to-week fluctuations observed in the raw data points, the model fits smooth linear trajectories that balance individual trends with the overall population average.

Figure 8. Residuals vs Fitted

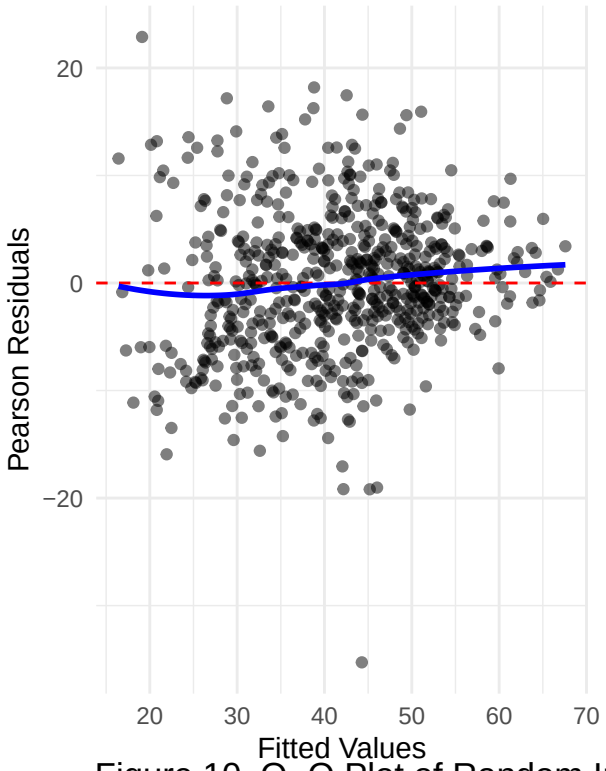


Figure 9. Residuals vs Time

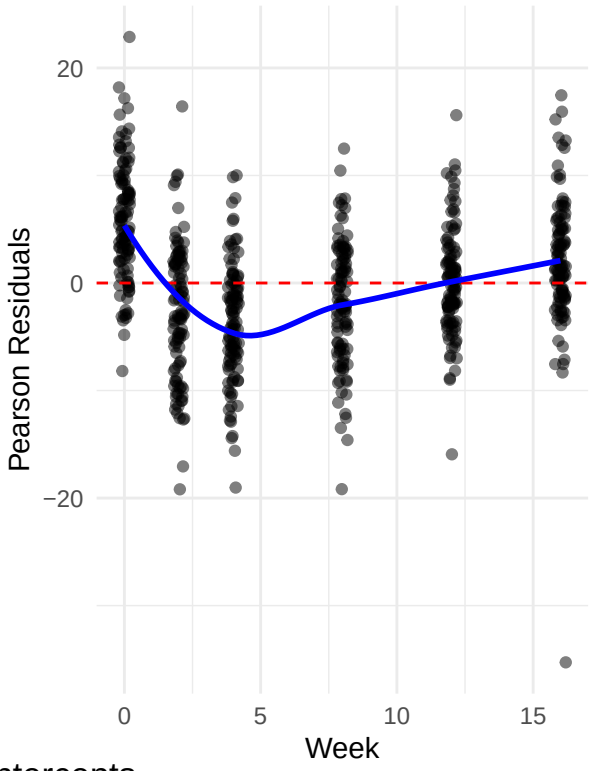
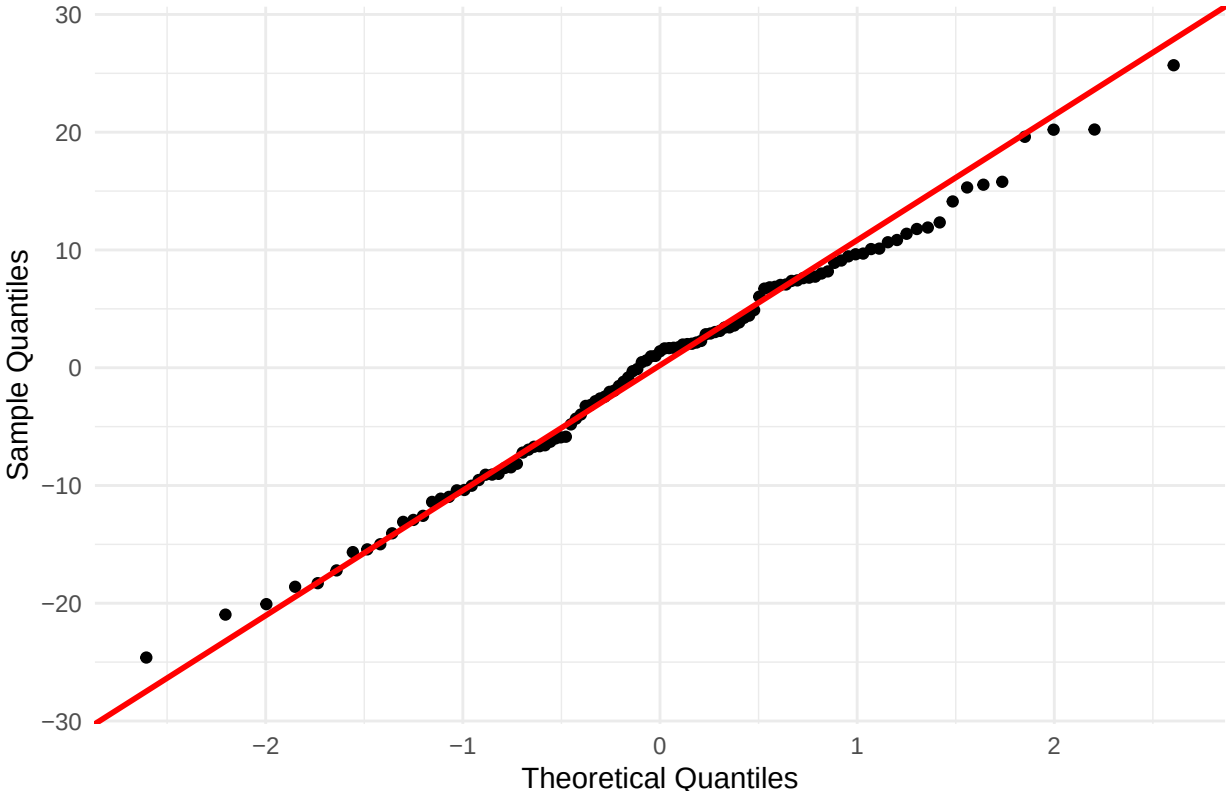


Figure 10. Q−Q Plot of Random Intercepts



The plot of Pearson residuals against fitted values (Figure 8) shows a random scatter of points around the zero line without a distinct funnel shape. This indicates that the assumption of constant variance

(homoscedasticity) is reasonably satisfied, and the linear mean structure is adequate. The plot of residuals against time (Figure 9) displays a slight non-linear trend. The dip suggests that the model is overpredicting the TWSTRS scores around Weeks 4 and 5. However, the residuals generally remain centered near zero, indicating the linear approximation is adequate for the primary analysis.

The Q-Q plot for the random intercepts (Figure 10) assesses the assumption that subject-specific deviations follow a normal distribution. The points fall closely along the red diagonal reference line, with only negligible deviations at the extreme tails. This confirms that the assumption of normally distributed random effects is well satisfied.

We also confirm that the model converged successfully.

Discussion

The actual trial did an ANCOVA model using week 4 TWSTRS score as the primary outcome measure, which does not account for treatment effects over time (Brashear et al., 1999). The best model of course depends on which research question an investigator wants answered. In this project, GLMM and GEE were the best choices for examining treatment effects over time. None of the models fit here showed a significant treatment effect.

References

Brashear, A., Lew, M. F., Dykstra, D. D., Comella, C. L., Factor, S. A., Rodnitzky, R. L., Trosch, R., Singer, C., Brin, M. F., Murray, J. J., Wallace, J. D., Willmer–Hulme, A., & Koller, M. (1999). Safety and efficacy of NeuroBloc (Botulinum toxin type b) in type A–responsive cervical dystonia. *Neurology*, *53*(7), 1439–1439. https://doi.org/10.1212/WNL.53.7.1439

Jankovic, J., Tsui, J., & Brin, M. F. (2023). Treatment of cervical dystonia with Botox (Onabotulinumtoxina): Development, insights, and impact. *Medicine*, *102*(S1), e32403. https://doi.org/10.1097/MD.0000000000032403

Wetmore, E., Roberts, H., Livinski, A. A., Camacho, T., Eaton, C., Norato, G., Hallett, M., & Stacy, M. (2025). Clinical response to placebo botulinum toxin injection in cervical dystonia—a systematic review and meta-analysis. *Dystonia*, *4*, 14297. https://doi.org/10.3389/dyst.2025.14297