# Examining BotB treatment effects over time for CD patients

Lily Bai Nathalie Blasco Yihao Peng Lauren Rackley Sirui Wu

2025-11-28

## 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

### *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. Prior to fitting models, sex and treatment group were converted to factor variables in R. Before fitting GEE and GLMM, a unique subject ID variable was created by combining site and ID. For GLM, age, sex, week, and treatment were used as covariates with male and placebo being the reference groups. For all three 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 4.5.2, R Core Team, 2025).

## Results

### *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. Prior participation in BotB trials precluded patients from enrolling in the study (Brashear et al., 1999). The study included 109 patients (67 (61%)) females. In the study, 36 patients were randomized to placebo, 36 were randomized to 5000U BotB and 37 were randomized to 10000U BotB. 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. Pearson's Chi Square test was used to compare sex differences among treatment groups. One-way

ANOVA was used to compare differences among treatment groups for TWSTRS total score at baseline and age (Table 1).

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) | | | | | 0.5195 |
| 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 | | | | | 0.2911 |
| 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; One-way analysis of means

**Generalized Linear Model (GLM)**

A generalized linear model (GLM) with a Gaussian link was fit to evaluate the relationship between TWSTRS scores and age, sex, week, treatment group, and the week-by-treatment interaction.

Table 2: GLM Model Summary (Gaussian Link)

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| (Intercept) | 38.836 | 2.730 | 14.227 | 0.000 |
| week | 0.067 | 0.156 | 0.429 | 0.668 |
| treat5000 U | -0.844 | 1.977 | -0.427 | 0.670 |
| treat10000 U | -2.929 | 1.962 | -1.493 | 0.136 |
| age | 0.017 | 0.042 | 0.407 | 0.684 |
| sexFemale | 2.147 | 1.070 | 2.006 | 0.045 |
| week:treat5000 U | 0.122 | 0.220 | 0.555 | 0.579 |
| week:treat10000 U | 0.374 | 0.219 | 1.708 | 0.088 |

The coefficient for sex was 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. Because a higher TWSTRS score indicates worse outcomes, this suggests that females exhibit worse symptom severity than males, which is consistent with the known epidemiology of cervical dystonia.

The interaction coefficients between week and treatment were positive, meaning that treated groups tended to show greater increases in TWSTRS over time compared with the control group. Since increases represent worsening symptoms, this indicates a trend toward deterioration over time in the treatment groups, although the interaction effects did not reach statistical significance.

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**

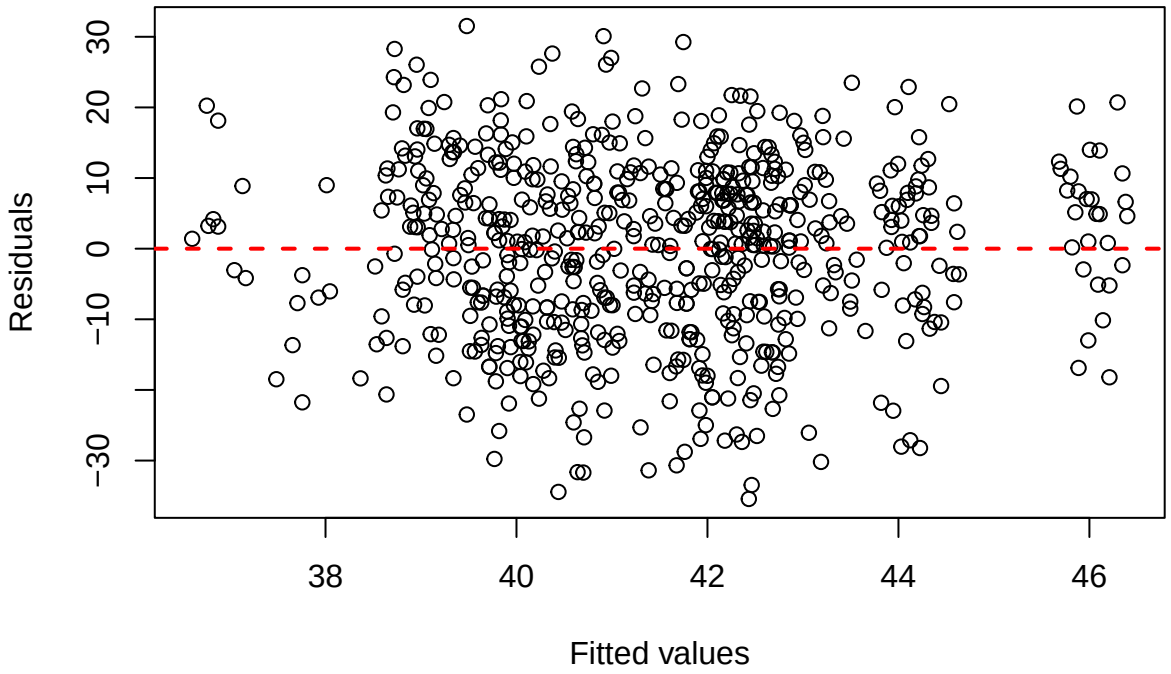



Figure 1 shows the residuals versus the fitted values. The absence of strong patterns or fanning suggests that the linearity and homoscedasticity assumptions are reasonably met.

**Figure 2: Q−Q Plot**

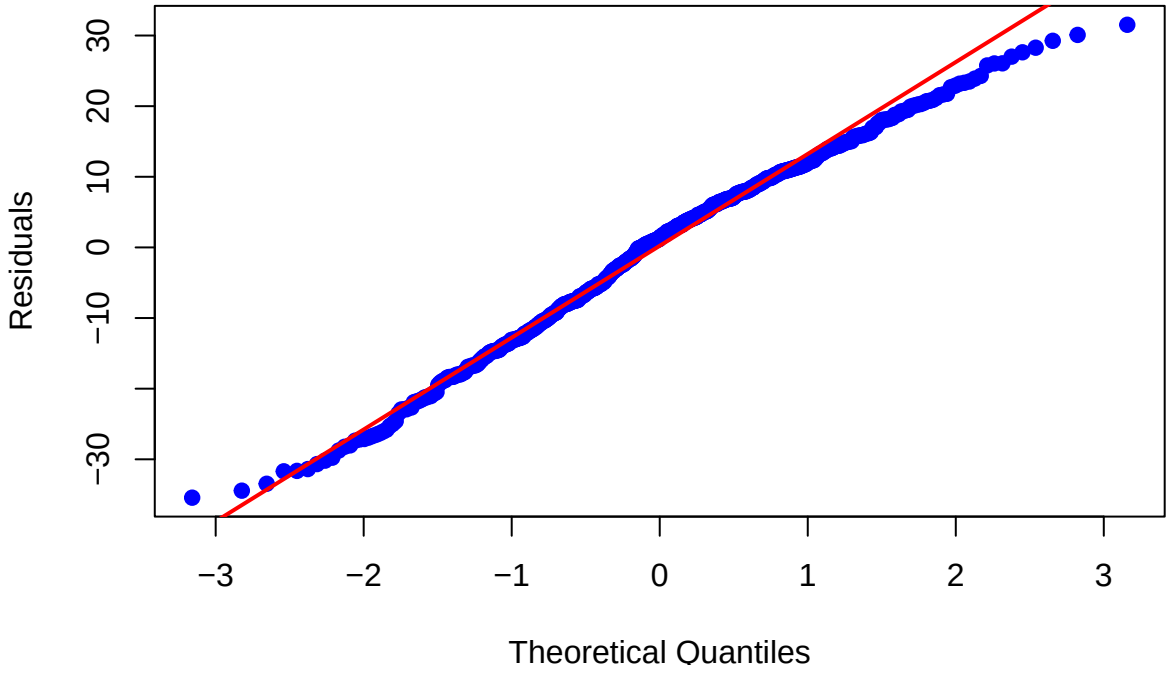
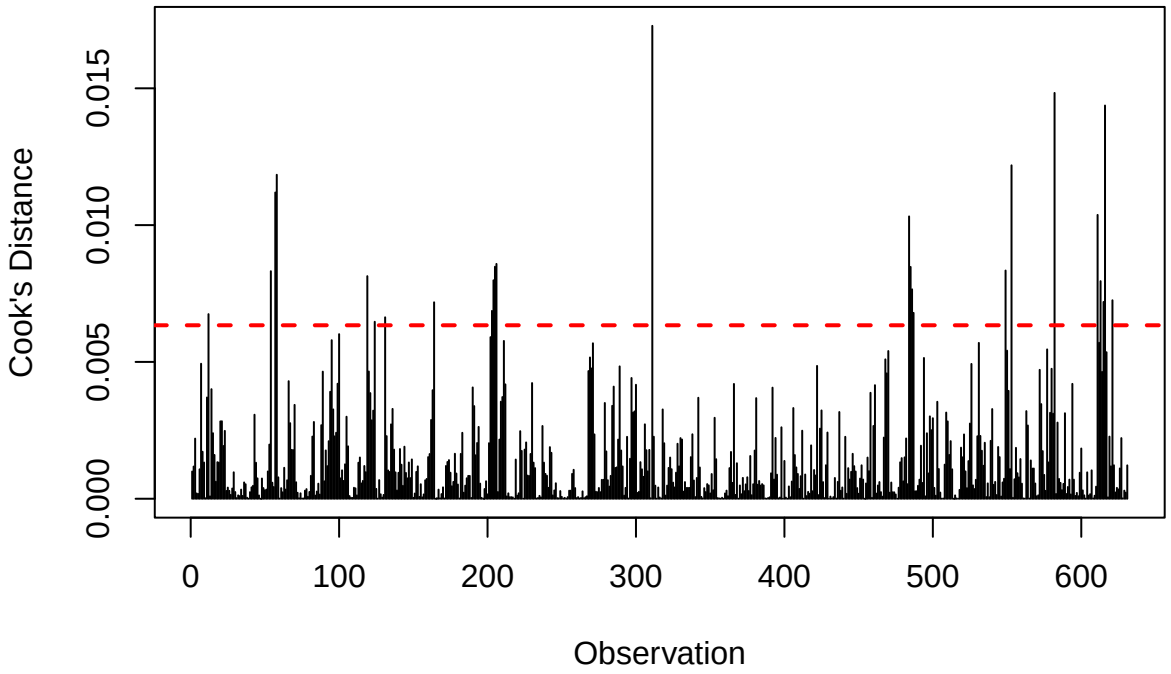


Figure 2 presents a QQ plot of the residuals, which largely follow the 45 degree reference line with mild deviations in the tails. This indicates that the normality assumption is approximately satisfied.

**Figure 3: Cook's Distance Plot**




Finally, figure 3 displays Cook's distance for all observations. The conventional threshold, 4/n, where n is the number of observations, 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) | 37.233 | 2.562 | 14.534 | 0.000 |
| week | 0.145 | 0.147 | 0.985 | 0.325 |
| treat5000 U | -0.486 | 1.847 | -0.263 | 0.792 |
| treat10000 U | -3.631 | 1.815 | -2.001 | 0.046 |
| age | 0.038 | 0.039 | 0.957 | 0.339 |
| sexFemale | 2.990 | 1.002 | 2.986 | 0.003 |
| week:treat5000 U | 0.082 | 0.209 | 0.392 | 0.695 |
| week:treat10000 U | 0.319 | 0.205 | 1.560 | 0.119 |

Excluding influential observations resulted in modest shifts in the parameter estimates and slightly improved precision. Notably, the effect of sex remained statistically significant. The treatment effect of 10,000 U at baseline also became stronger and statistically significant while the interaction coeffecients shrunk closer to 0. 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.

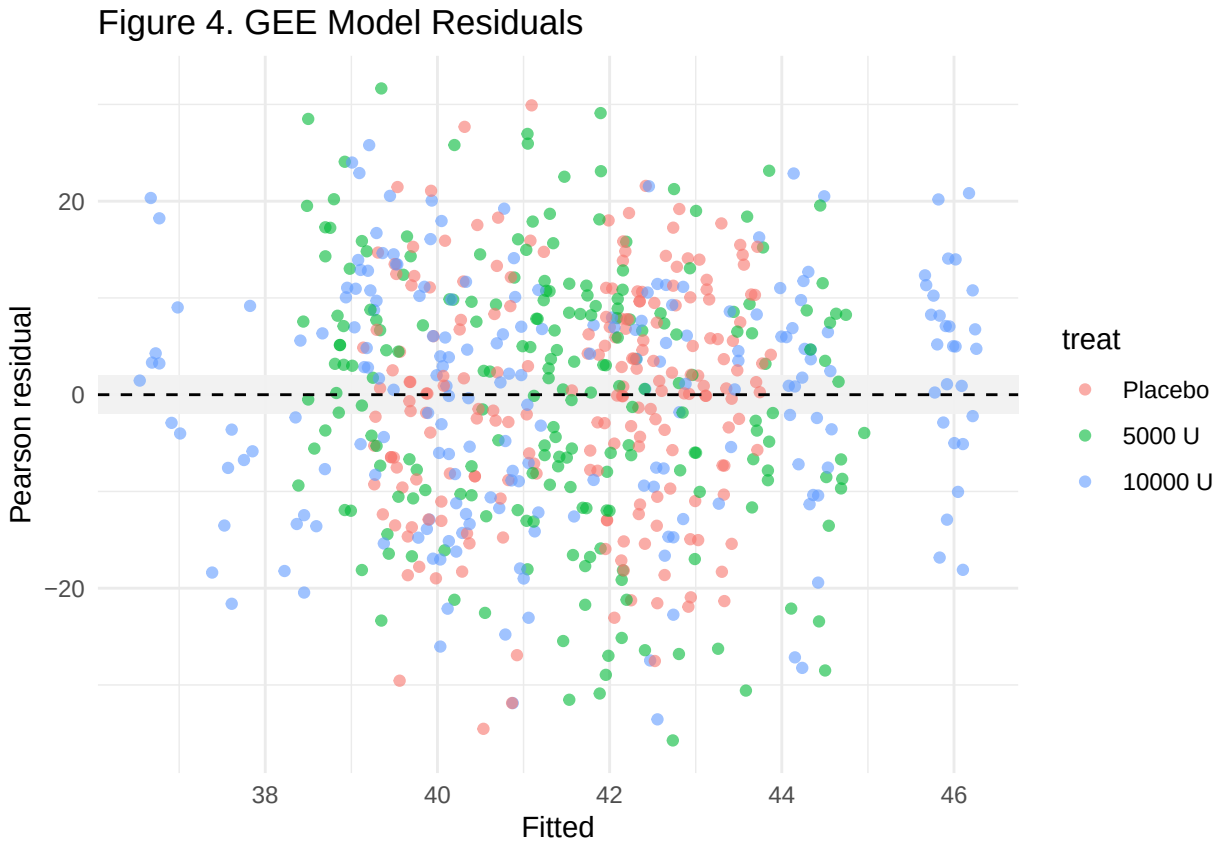



Figure 4 shows Pearson residuals plotted against the fitted TWSTRS values with points colored by treatment group. The residuals are centered around zero and do not show a strong funnel shape, supporting the constant-variance assumption. However, many residuals fall outside $\pm 2$, indicating wide individual 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.

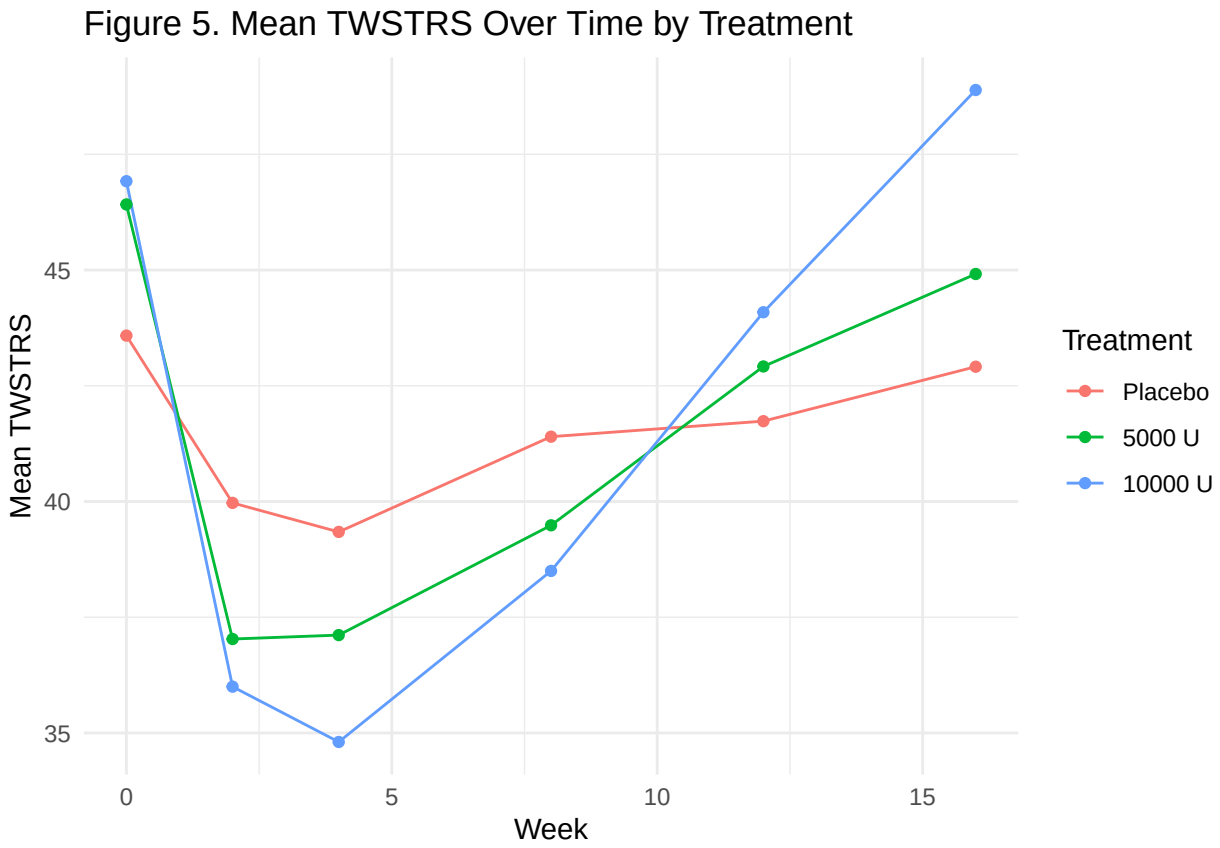



Figure 5 shows mean TWSTRS over time by treatment group. All groups show decreases from baseline through weeks 2–4, suggesting early improvement, followed by increases after week 4. Around week 10, both active treatment groups begin to show higher TWSTRS scores than the placebo group.

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

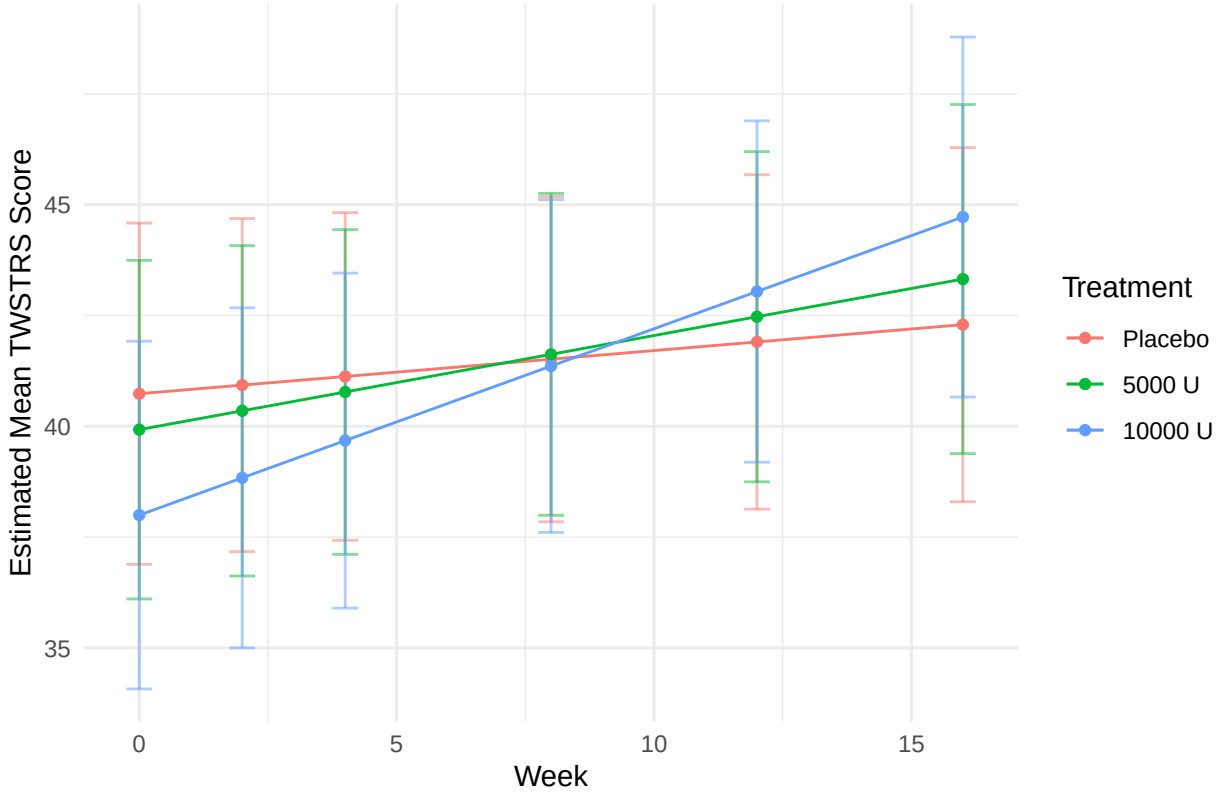



Figure 6 illustrates the model-predicted trajectories for TWSTRS across the three treatment groups. The blue line representing 10000 U treatment and the red line for placebo cross each other, showing the significant interaction term between 10000 U and week. The relationship between treatment and score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

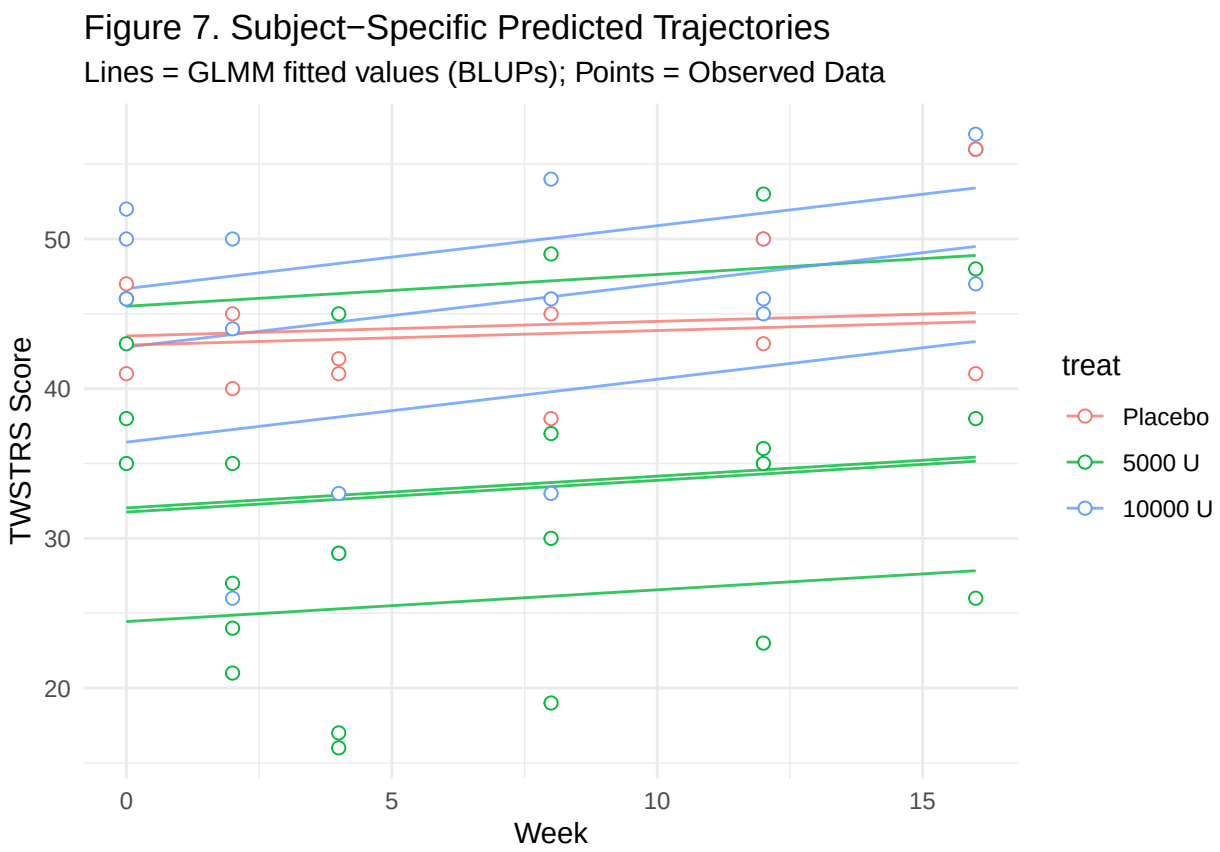



Figure 7 displays the observed TWSTRS scores (points) overlaying the GLMM fitted trajectories (lines) for a random sample of 12 subjects. The heterogeneity at baseline confirms the necessity of including a random intercept for each subject to account for patient-specific heterogeneity, suggesting that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

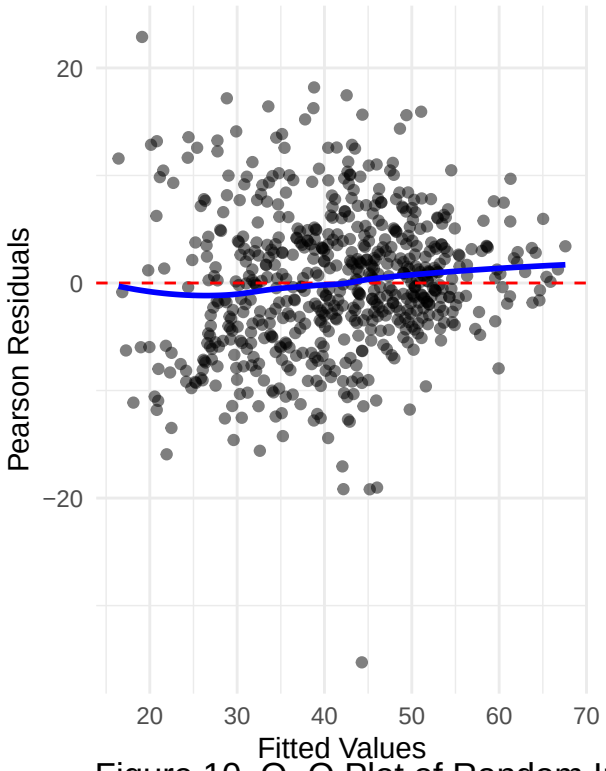



Figure 9. Residuals vs Time

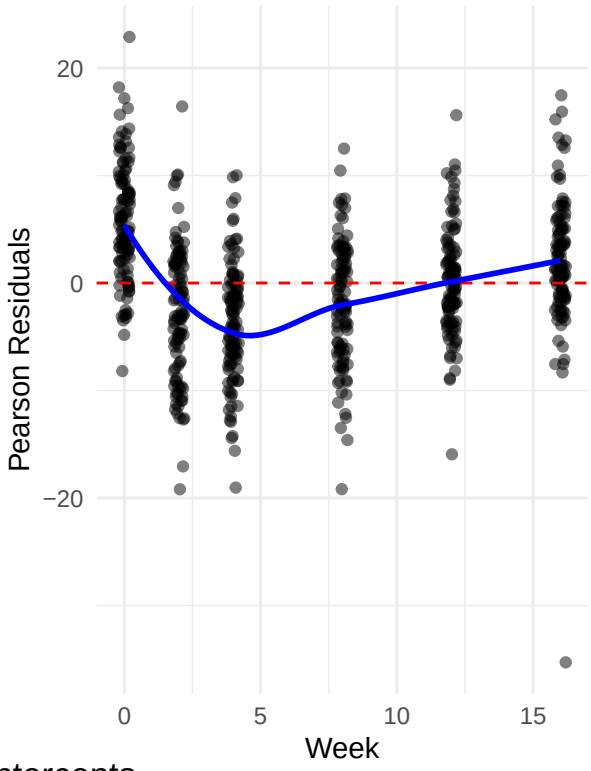
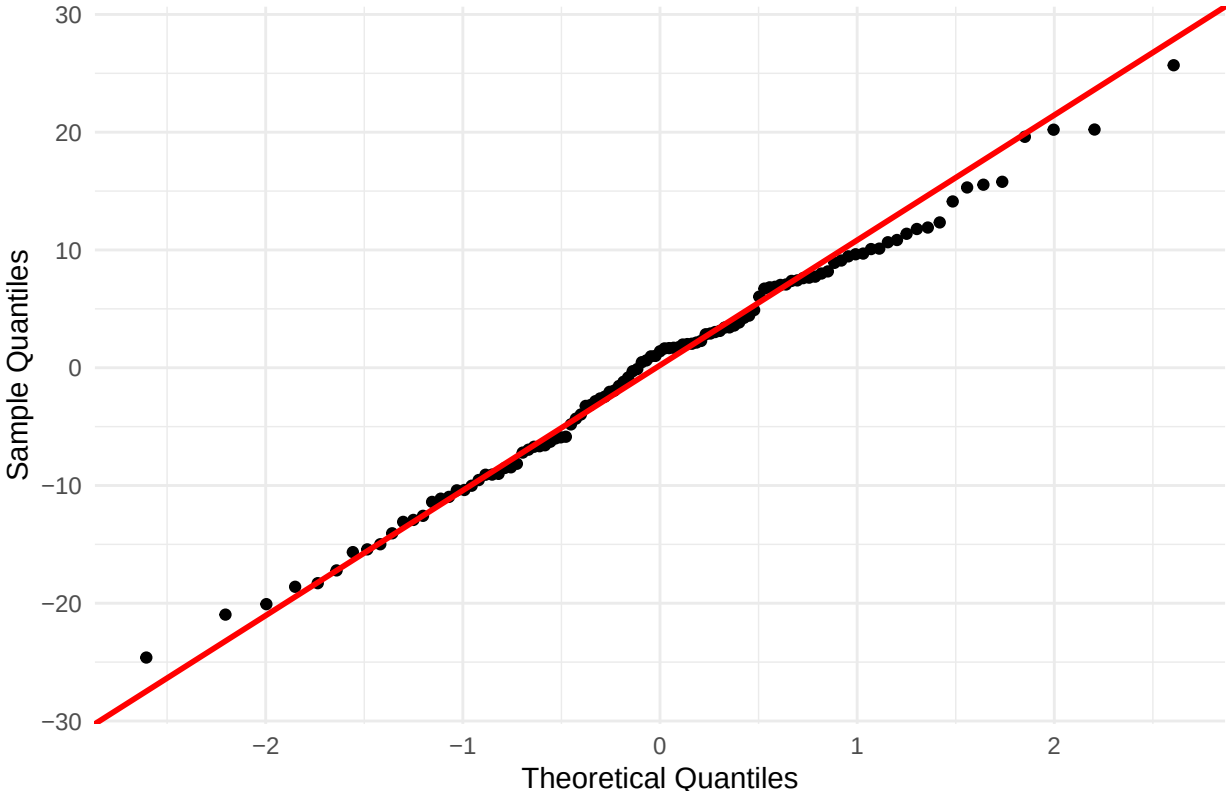



Figure 10. Q−Q Plot of Random Intercepts




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

This study investigated longitudinal changes in cervical dystonia severity following administration of botulinum toxin type B (BotB) using data from a randomized clinical trial with three treatment arms: placebo, 5000U, and 10000U. TWSTRS scores were collected at multiple time points over a 16-week follow-up period, providing a repeated-measures framework for evaluating differences in symptom trajectories across treatment groups.

As an initial exploratory step, a generalized linear model (GLM) was fit to examine associations between treatment, time, and relevant demographic variables. However, because the GLM assumes independence of observations and therefore ignores the within-subject correlation inherent to repeated measures, its coefficient estimates cannot be interpreted as valid inferential quantities. The GLM served primarily as a descriptive benchmark, motivating the use of methods that explicitly account for the covariance structure of longitudinal data. Accordingly, inference was based on two widely used frameworks for repeated-measures analysis: generalized estimating equations (GEE) and generalized linear mixed models (GLMM).

Both GEE and GLMM yielded highly concordant fixed-effect estimates. At baseline, neither active treatment group differed significantly from placebo, and neither age nor sex exhibited meaningful associations with initial severity or subsequent change. Across both modeling approaches, the 5000U group showed no statistically detectable deviation from the placebo trajectory, as evidenced by a nonsignificant treatment-by-time interaction. In contrast, the 10000U group consistently demonstrated a significantly more positive slope relative to placebo, indicating a faster rate of symptom worsening over the 16-week period. Given that higher TWSTRS scores correspond to greater clinical impairment, this finding suggests that the higher dose may be associated with an adverse rather than therapeutic longitudinal profile. The consistency of this effect across both the marginal (GEE) and conditional (GLMM) modeling frameworks strengthens the robustness of the conclusion.

Although the estimated treatment effects were similar across GEE and GLMM, the two approaches target different estimands. GEE yields population-averaged (marginal) effects, characterizing the mean response pattern in the study population. GLMM, through inclusion of random effects, generates subject-specific (conditional) estimates and partitions variation into between-subject and within-subject components. Model comparison indicated that a random-intercept structure was sufficient to capture substantial heterogeneity in baseline TWSTRS levels, whereas including random slopes did not improve model fit, implying limited between-subject variability in rates of change. Diagnostic evaluations supported the adequacy of the mixed-model assumptions: random intercepts were approximately normally distributed, fitted-versus-residual plots exhibited no major departures from homoscedasticity, and residual-versus-time plots showed only mild non-linear patterns.

Because GEE and GLMM address different scientific aims, the choice between them depends on whether inference is sought for marginal population-level effects (favoring GEE) or for individual-level trajectories and conditional effects (favoring GLMM). In this analysis, both approaches converged on the same substantive conclusion: neither the 5000U nor the 10000U dose of BotB produced improvement relative to placebo, and the higher dose was associated with a significantly more pronounced deterioration in symptoms over time.

Several limitations should be considered. First, modeling time as a linear term simplifies an empirically nonlinear trajectory characterized by early improvement followed by subsequent worsening. More flexible

time-varying specifications, such as spline-based or piecewise models, may offer a more accurate representation of the underlying mean structure. Second, the sample size within each treatment arm was modest, limiting statistical power, particularly for detecting smaller interaction effects or subtle differences in longitudinal profiles. Third, the dataset lacked detailed clinical covariates that could potentially account for heterogeneity in treatment response across individuals. Future analyses incorporating richer baseline information, nonlinear longitudinal modeling, and more frequent early follow-up assessments would enable a more comprehensive characterization of dose–response dynamics in cervical dystonia.

## 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