Gait Analysis of Children with Cerebral Palsy

Kevin Nguyen

October 24, 2022

1 Executive Summary

The most common cause of motor impairment in children is cerebral palsy (CP). Individuals with CP have complicated motor disorders because they have primary deficiencies, such as muscle spasticity and muscle weakness, and secondary deficiencies, like muscle contractures and bone deformities. The main dysfunctions are motor impairments that impair mobility and posture, making walking difficult. The data set provided by the client consists of measurements on children with varying types and severity of CP. The overall goal is to use this information to assess if a clinical study can be conducted to evaluate the efficacy of a strength intervention to increase walking speed.

This report appropriately addresses all three of the client's questions. First, there is likely to be a weak positive relationship between strength and walking speed. Second, it is likely the magnitude of the association between strength and walking speed depends on the type of CP but not severity. Third, there is likely to be a weak positive relationship between adequate walking speed and strength. Furthermore, it is unlikely the magnitude of the association between adequate walking speed and strength depends on the type of CP or severity.

2 Introduction

2.1 General Background

The most common cause of motor impairment in children is cerebral palsy (CP). Individuals with CP have complicated motor disorders because they have primary deficiencies, such as muscle spasticity and muscle weakness, and secondary deficiencies, like muscle contractures and bone deformities. The main dysfunctions are motor impairments that impair mobility and posture, making walking difficult. Walking is essential for daily activities and social participation. Clinical gait analysis is frequently employed to identify, characterize, and evaluate the deficiencies of patients with CP and is incorporated in clinical decision-making for patients with complicated gait disorders.

The data set provided by the client consists of measurements on children with varying types and severity of CP. The overall goal is to use this information to assess if a clinical study can be conducted to evaluate the efficacy of a strength intervention to increase walking speed. The relationship between an overall measure of bodily strength and walking speed is of particular interest.

2.2 Objectives

This consultation's main objective is to provide answers to the following three questions:

1. **Primary Question**: Is there a relationship between strength and walking speed, with or without adjustment for height, weight or BMI?

2. **Secondary Question 1**: Does the strength of this association depend on severity or type of CP?

3. **Secondary Question 2**: Generally, a walking speed of 80 cm/sec is considered adequate for community dwelling. Is there an association between adequate walking speed and strength, and does this association depend on the type and severity of CP?

2.3 Data Description

The client's raw data is composed of 322 observations and 11 columns. The data set contains the following variables:

- **ID**: The subject ID.

- **GMFSCLevel**: A measure of the functional ability of the child using the Gross Motor Function Classification System. This variable is ordinal, consisting of values ranging from 1 to 3. Higher values indicate increasing levels of severity.

- **CPtype**: The type of CP. Values in this variable include Diplegia (CP affects both sides of the body) or Hemiplegia (CP affects one side of the body).

- **Involved_Side**: The side of the body CP affects. Values in this variable include Bilateral, Left, or Right.

- **Gender**: The gender of the child. Values in this variable include Male or Female.
- **Age**: The age of the child in years.
- **Standing˙Height**: The standing height of the child in centimeters.
- **Weight**: The weight of the child in kilograms.
- **BMI**: The body mass index (BMI) of the child. The formula for BMI is weight (in kilograms) divided by height (in meters) squared.
- **Walking_Speed**: The walking speed of the child in centimeters per second.
- **TotStrengthSum**: A summary measure of total strength.

3 Project Approach

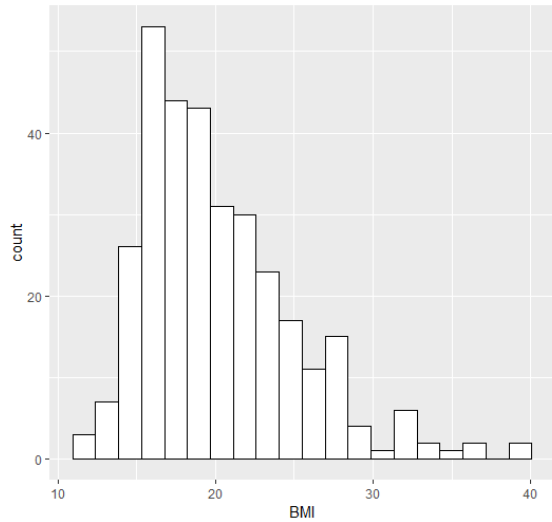
3.1 Data Cleaning

Processing the raw data involved the following four actions:

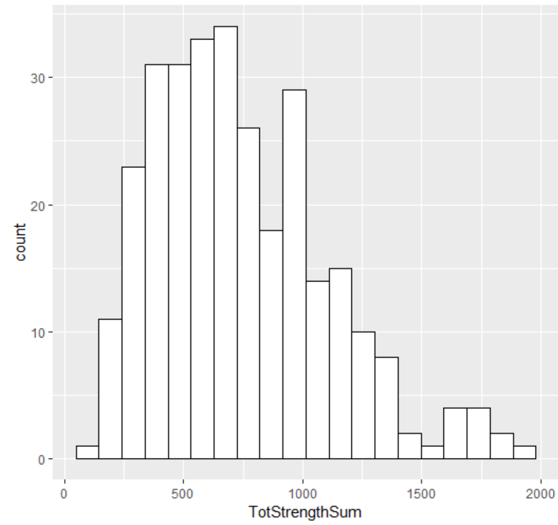
1. For BMI and TotStrengthSum, missing values were imputed with with the median.
2. For Walking_Speed, missing values were imputed with the mean.
3. Walking_Speed was converted to a binary variable to create a new variable named WS80.
4. GMFCSLevel, CPtype, Involved_Side, Gender, and WS80 were converted to a factor data type.

3.1.1 Imputation of Median in BMI and TotStrengthSum

The data set contains 1 missing value for BMI and 7 missing values for TotStrengthSum. We examined the distribution for BMI and TotStrengthSum to determine whether to impute the mean or median for missing values. Figure 1 shows both variables have a right-skewed distribution. Therefore, we opted to impute the median for the missing values.



(a) BMI



(b) TotStrengthSum

Figure 1: Distributions of BMI and TotStrengthSum

3.1.2 Imputation of Mean in Walking_Speed

The data set contains 24 missing values for Walking_Speed. We examined the distribution for Walking_Speed to determine whether to impute the mean or median for missing values. Figure 2 shows Walking_Speed has a fairly symmetric distribution. Therefore, we opted to impute the mean for the missing values.

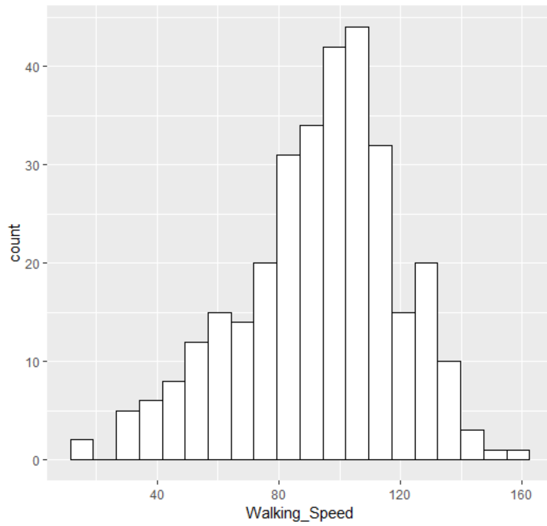


Figure 2: Distribution of Walking_Speed

3.1.3 Conversion to Binary Variable for Walking_Speed

To apply logistic regression, the response variable should be binary. Therefore, by using the `ifelse()` function, we created a new variable named WS80 that converts Walking_Speed to a binary variable. A value is labeled as 1 if Walking_Speed ≥ 80 and 0 otherwise.

3.1.4 Conversion to Factor Data Type for GMFCSLevel, CPtype, Involved_Side, Gender, and WS80

GMFCSLevel, CPtype, Involved_Side, Gender, and WS80 were converted to a factor data type using the `as.factor()` function because we wish to use them as categorical variables.

3.2 Methodology

We implemented linear regression to answer the **Primary Question** and **Secondary Question 1**. The value of a variable can be predicted using linear regression based on the values of the other variables. By using one or more independent variables that best predicts the value of the response variable, linear regression estimates the coefficients of the linear equation. Linear regression fits a straight line that minimizes the discrepancies between predicted and actual output values.

We implemented logistic regression, a supervised machine learning model, to answer **Secondary Question 2**. Based on a given data set of independent variables, logistic regression calculates the likelihood that an event will occur. A logit transformation is applied on the odds, the probability of success divided by the probability of failure, in logistic regression. The transformed odds are often referred to as log odds. In this model, the beta parameters, or coefficients, are often estimated via maximum likelihood estimation. This algorithm evaluates various beta values over a number of iterations to get the best match for the log odds. To determine the most accurate parameter estimate, logistic regression aims to maximize the log likelihood function, which is produced by all of these iterations. Following the identification of the best coefficients, the conditional probabilities for each observation can be computed, logged, and summed to provide a predicted probability.

4 Results

First, we checked for outliers by creating boxplots for each quantitative variable. Figure 3 illustrates there are outliers present in Weight, BMI, Walking_Speed, and TotStrengthSum.

Despite the presence of outliers, we do not want to discard any of our client's data due to the time and effort the children and researchers put into the clinical trial. We should therefore be mindful of the existence of these outliers moving forward.

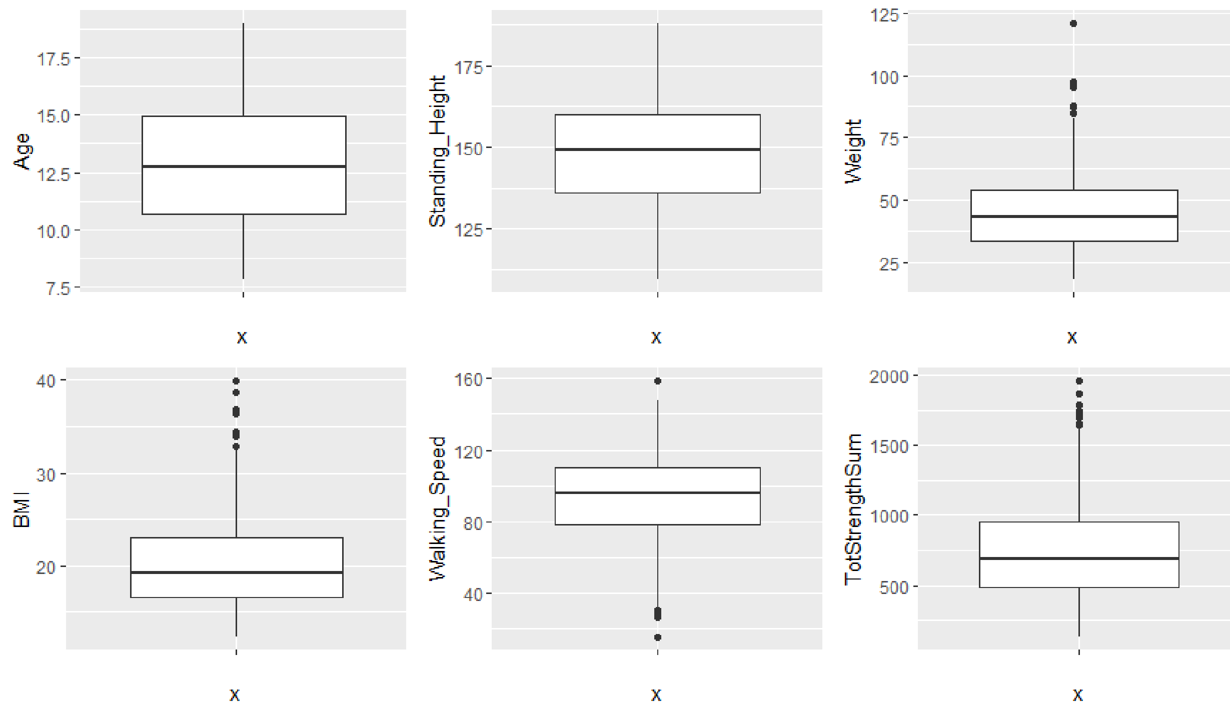


Figure 3: Boxplots of Quantitative Variables

Second, we fitted an unadjusted linear regression model, with Walking_Speed as the response, using the `lm()` function:

$$\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 TotStrengthSum$$

Using the `summary()` function, we estimated the coefficients for the unadjusted linear regression model. Figure 4 shows TotStrengthSum is statistically significant at $\alpha = 0.05$. Furthermore, the estimated coefficient of 0.026 for TotStrengthSum in Figure 4 indicates there is a weak positive relationship between strength and walking speed.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	73.635	3.129	23.53	0
TotStrengthSum	0.026	0.004	6.84	0

Figure 4: Estimated Coefficients for $\hat{y}_1$

Third, we fitted an adjusted linear regression model, with Walking_Speed as the response, using the `lm()` function:

$$\hat{y}_2 = \hat{\beta}_0 + \hat{\beta}_1 TotStrengthSum + \hat{\beta}_2 Standing\_Height + \hat{\beta}_3 Weight + \hat{\beta}_4 BMI$$

Using the `summary()` function, we estimated the coefficients for the adjusted linear regression model. Figure 5 shows TotStrengthSum is statistically significant at $\alpha = 0.05$. Furthermore, the estimated coefficient of 0.024 for TotStrengthSum in Figure 5 indicates there is a weak positive relationship between strength and walking speed.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	56.787	49.452	1.148	0.252
TotStrengthSum	0.024	0.004	5.613	0.000
Standing_Height	0.218	0.331	0.659	0.510
Weight	-0.034	0.546	-0.062	0.950
BMI	-0.613	1.283	-0.478	0.633

Figure 5: Estimated Coefficients for $\hat{y}_2$

Fourth, we fitted the following two linear regression models, with Walking_Speed as the response, using the `lm()` function:

1. $\hat{y}_3 = \hat{\beta}_0 + \hat{\beta}_1 TotStrengthSum + \hat{\beta}_2 GMFCSLevel + \hat{\beta}_3 TotStrengthSum : GMFCSLevel$

2. $\hat{y}_4 = \hat{\beta}_0 + \hat{\beta}_1 TotStrengthSum + \hat{\beta}_2 CPtype + \hat{\beta}_3 TotStrengthSum : CPtype$

Note: TotStrengthSum:GMFCSLevel represents the interaction term between TotStrengthSum and GMFCSLevel. Likewise, TotStrengthSum:CPtype represents the interaction term between TotStrengthSum and CPtype.

Using the `summary()` function, we estimated the coefficients for both linear regression models. Figure 6 shows the interaction between TotStrengthSum and GMFCSLevel is not statistically significant at $\alpha = 0.05$. However, Figure 7 shows the interaction between TotStrengthSum and CPtype is statistically significant at $\alpha = 0.05$.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	103.507	4.174	24.797	0.000
TotStrengthSum	0.004	0.004	0.932	0.352
GMFCSLevel2	-19.871	6.178	-3.216	0.001
GMFCSLevel3	-53.515	6.831	-7.834	0.000
TotStrengthSum:GMFCSLevel2	0.011	0.008	1.406	0.161
TotStrengthSum:GMFCSLevel3	0.014	0.010	1.338	0.182

Figure 6: Estimated Coefficients for $\hat{y}_3$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	63.985	3.603	17.759	0.000
TotStrengthSum	0.034	0.005	7.251	0.000
CPtypeHemiplegia (left)	40.067	8.481	4.725	0.000
CPtypeHemiplegia (right)	33.618	8.413	3.996	0.000
TotStrengthSum:CPtypeHemiplegia (left)	-0.031	0.010	-3.164	0.002
TotStrengthSum:CPtypeHemiplegia (right)	-0.025	0.010	-2.631	0.009

Figure 7: Estimated Coefficients for $\hat{y}_4$

Lastly, we fitted the following three logistic regression models, with WS80 as the response, using the `glm()` function with the argument `family = "binomial"`:

1. $\hat{y}_5 = \hat{\beta}_0 + \hat{\beta}_1 TotStrengthSum$
2. $\hat{y}_6 = \hat{\beta}_0 + \hat{\beta}_1 TotStrengthSum + \hat{\beta}_2 GMFCSLevel + \hat{\beta}_3 TotStrengthSum : GMFCSLevel$
3. $\hat{y}_7 = \hat{\beta}_0 + \hat{\beta}_1 TotStrengthSum + \hat{\beta}_2 CPtype + \hat{\beta}_3 TotStrengthSum : CPtype$

Using the `summary()` function, we estimated the coefficients for the three logistic regression models. Figure 8 shows TotStrengthSum is statistically significant at $\alpha = 0.05$. Furthermore, the estimated coefficient of 0.003 for TotStrengthSum in Figure 8 indicates there is a weak positive relationship between adequate walking speed and strength. Figure 9 shows the interaction between TotStrength and CPtype is not statistically significant at $\alpha = 0.05$. Similarly, Figure 10 shows the interaction between TotStrength and GMFCSLevel is not statistically significant at $\alpha = 0.05$.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.687	0.335	-2.050	0.04
TotStrengthSum	0.003	0.001	5.178	0.00

Figure 8: Estimated Coefficients for $\hat{y}_5$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.007	0.378	-2.667	0.008
TotStrengthSum	0.003	0.001	4.551	0.000
CPtypeHemiplegia (left)	3.132	1.808	1.732	0.083
CPtypeHemiplegia (right)	0.809	1.424	0.568	0.570
TotStrengthSum:CPtypeHemiplegia (left)	-0.001	0.002	-0.502	0.616
TotStrengthSum:CPtypeHemiplegia (right)	0.002	0.003	0.610	0.542

Figure 9: Estimated Coefficients for $\hat{y}_6$

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.433	0.881	1.627	0.104
TotStrengthSum	0.002	0.001	1.362	0.173
GMFCSLevel2	-0.031	1.068	-0.029	0.977
GMFCSLevel3	-5.049	1.307	-3.863	0.000
TotStrengthSum:GMFCSLevel2	-0.002	0.001	-1.087	0.277
TotStrengthSum:GMFCSLevel3	0.002	0.002	1.117	0.264

Figure 10: Estimated Coefficients for $\hat{y}_7$

5 Conclusions

In this report, we feel we have appropriately addressed all three of the client's questions. In response to the **Primary Question**, our analysis found TotStrengthSum to be statistically significant at $\alpha = 0.05$ for the unadjusted and adjusted linear regression models. Additionally, the estimated coefficient for TotStrengthSum from the unadjusted and adjusted linear regression models are 0.026 and 0.024, respectively, which suggests there is likely to be a weak positive relationship between strength and walking speed.

In response to **Secondary Question 1**, our analysis found the interaction between TotStrengthSum and GMFCSLevel to not be statistically significant at $\alpha = 0.05$. But, our analysis also found the interaction between TotStrengthSum and CPtype to be statistically significant at $\alpha = 0.05$. Therefore, it is likely the magnitude of the association between strength and walking speed depends on the type of CP but not severity.

In response to **Secondary Question 2**, our analysis found TotStrengthSum to be statistically significant at $\alpha = 0.05$ for the first logistic regression model. Moreover, the estimated coefficient for the first logistic regression model is 0.003, which suggests there is likely to be a weak positive relationship between adequate walking speed and strength. Also, our analysis found the interaction between TotStrength and CPtype nor the interaction between TotStrength and GMFCSLevel to not be statistically significant at $\alpha = 0.05$ for our second and third logistic regression models. Therefore, it is unlikely the magnitude of the association between adequate walking speed and strength depends on the type of CP or severity.