STA 550 Project: Statistical Dynamics of PROMs in CVT: A Longitudinal Analysis

Wakeel Kasali

April 5, 2024

Contents

Abstract

This study aims to develop a statistical model to uncover the complex interplay between symptom progression and health outcomes in patients with cerebral venous thrombosis (CVT). Focusing on linear mixed-effect models (LMMs) and generalized estimating equations (GEEs) with four covariates Headache, Nausea/Vomiting, Other symptoms, Total years of Education, the model were fitted using data from the national clinical trial and parallel registry for people with CVT collected over three years with 53 patients in the trial and another 50 in the registry.

The accuracy of the models performance was accessed using the Q-Q plots for the assumed normality distribution of the residuals. The model showed that Nausea/Vomiting, Other symptoms, Total years of Education were significant predictors of quality of life, while headache is not. The model's accuracy is based upon non-stringency of residuals normality assumption. The main limitation of the model was the potential for working better under data missing completely at random.

Our Cox proportional model provides an effective tool to identify whether delayed diagnosis contributes to worse functional or psychological outcomes. The GEE model suggests that the number of years invested on education, nausea/vomiting, and some other symptoms which are not very common among individual patients potentially have bad effect on quality of life. Further studies are needed to confirm the association and to test the model in different populations and settings.

1 Introduction

Cerebral venous thrombosis (CVT) is a rare type of stroke primarily affecting young women, often linked to oral contraceptive pills for birth control and childbirth. While outcomes appear positive with most patients regaining independence, lingering issues like headaches, fatigue, and depression are prevalent. This study investigates these discrepancies using patient-reported outcome measures (PROMs) to track symptom changes over a year for a better understanding of CVT's long-term effects which would pave the way for better patient care strategies.

The statistical questions of interest are:

- Whether the PROMs, capturing dimensions such as headache, mood,

fatigue, cognition, and quality of life, change over the course of time in individuals diagnosed with CVT.

- If there exists a pattern of interdependency among the various PROMs that could shed light on the interconnected nature of symptom domains in patients with CVT.

- Whether the timing of a cerebral venous thrombosis diagnosis impacts the progression of patient-reported health and psychological states within the year post-diagnosis.

The findings from this research are expected to enhance the comprehension of the longitudinal impacts of CVT, thereby contributing to the development of more effective patient management approaches.

2 Data description and Summaries

This is an experimental study that spanned three years and included a comprehensive follow-up period of 12 months for each participant. A total of 103 patients were involved, with 53 enrolled in the clinical trial and 50 included in the registry. Through the use of patient-reported outcome measures (PROMs) and neuroimaging, the study consistently assessed the longitudinal effects of cerebral venous thrombosis (CVT) on patients.

Fundamental to the study were outcome variables such as the modified Rankin Scale (mRS), Euro-QoL-5D (EQ-5D-5L), and the Visual Analog Scale (EQ-VAS), which served as primary instruments to evaluate the repercussions of CVT on patient well-being and recovery progression. These measures span ordinal, interval, and continuous data types, respectively.

The study also incorporated demographic and clinical variables. Each patient was assigned a unique identifier (ID), with additional continuous variables including Age, and time-related variables such as 'Time from symptoms to enrolment' and 'Symptoms to diagnosis'- measured in days. Sex/Gender and Ethnicity were categorized as categorical variables.

Clinical assessments and symptoms were well documented, capturing both the presence and absence of various symptoms like headache and nausea/vomiting (NV)- represented as binary categorical variables. Further continuous and categorical variables derived from clinical assessments included the NIHSS score, and neuroimaging findings such as venous infarct, midline

shift, and hemorrhage. Treatment specifics, such as the type of anticoagulant medication used (Anticoagulant) and the recanalization status at various intervals, were also categorized.

Patient-reported outcome measures (PROMs) included assessments over multiple time points for headache, mood, fatigue, cognition, and quality of life recorded as continuous variables. This also encompassed EQ5D and EQ-VAS scores for health-related quality of life, and scales such as FAS, HIT6, and PHQ9 for fatigue, headache impact, and depression severity. Notably, the dataset exhibited missing values, particularly within cognitive data, compared to other outcomes. This presents a challenge often encountered in longitudinal studies and recognizes the importance of considering flexible but cautious approach for handling missing data.

3 Missing data and lost-to-follow-up

In this randomized clinical trial, prior to analysis, a thorough assessment of data missingness is essential. Visualization tools, such as the one depicted in Figure 1, can brief the proportion of missing values across some outcomes within the dataset. Conventionally, missing data is deemed negligible when its prevalence falls below 5%. Under these conditions, analyses utilizing solely observed data are considered sufficient, provided that the extent and potential limitations due to missingness are clearly reported.

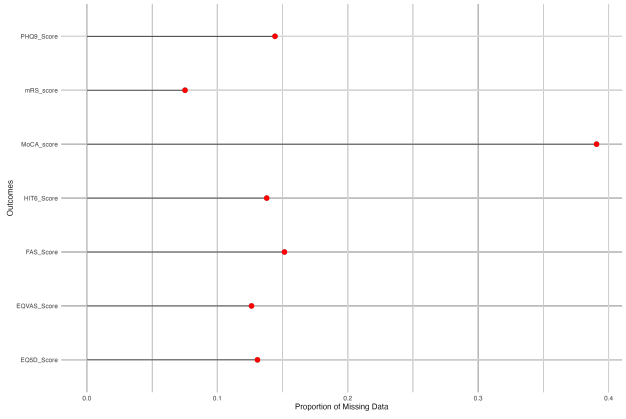

Figure 1: Proportion of missing data for outcomes scores

In instances where missing data surpasses the 5% threshold, but stays

below 40%, and is not solely concentrated in outcome variables, the use of multiple imputation methods offers a safe alternative. This approach assumes the data is missing at random - that is, missingness can be predicted by other variables within the dataset.

Despite the benefits of multiple imputation, trade-offs exist, particularly in terms of reduced statistical power and potential bias in analyses restricted to observed data. It is noteworthy that currently available commercial methods cater exclusively to continuous variables. Figure 2 presents the proportion of missing data within outcome variables, indicating a range between 5% to 40%, which may warrant the application of multiple imputation.

Furthermore, the cognitive variable (MoCA score) manifests significant data absence. To understand the lost rates and their behaviour with time, the Kaplan-Meier curve, as shown in Figure 2, is relevantly descriptive. It reveals high survival probabilities for the majority of the study, with a marked decrease towards the end, indicating a specific issue with MoCA assessments in later stages of the study.

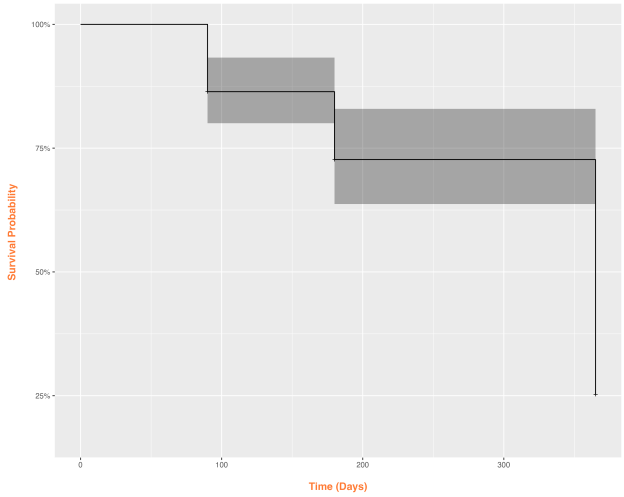

Figure 2: Time-to-Event Analysis of MoCA Scores: A Kaplan-Meier Survival Curve

4 Exploratory Data Analysis

In the analysis of cerebral venous thrombosis (CVT) outcomes, a multidimensional approach to health assessment is crucial. The exploratory data analysis, focusing on various health dimensions, aims to avoid redundancy through the use of correlated measures. Figure 3 illustrates the interrelationships among CVT outcomes, such as the depression module (PHQ9_Score), health-related quality of life (EQ5D_Score), EuroQol Visual Analogue Scale (EQVAS_Score), Headache Impact Test-6 (HIT6_Score), Fatigue Assessment Scale (FAS_Score), and the degree of disability or dependence (mRS_Score). Notably, these measures do not demonstrate strong correlations with each other as it is confirmed by the pearson correlation.

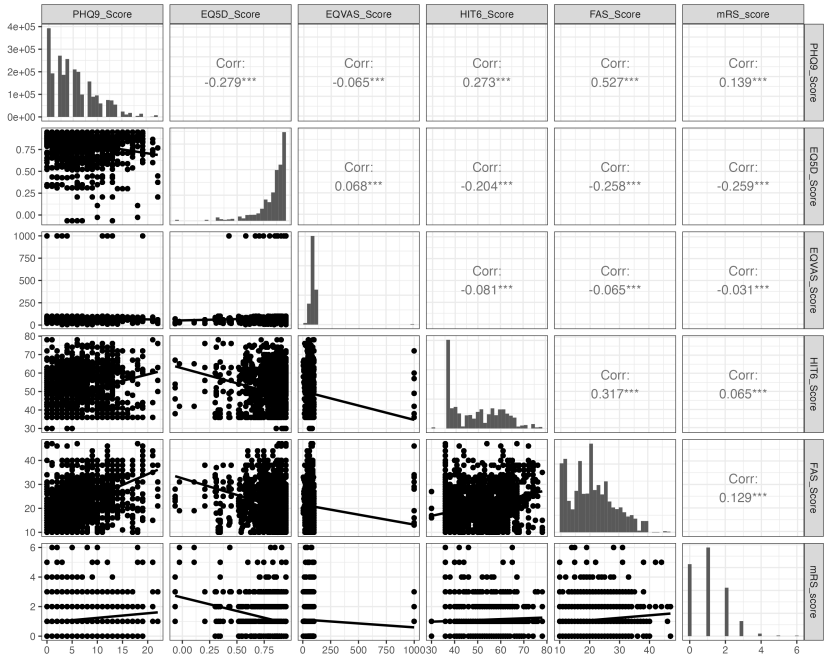

Figure 3: Interdependency of health outcome measures

The EQ-5D score, a key measure within this study, which connects various health domains, including mobility, self-care, and mental health aspects like anxiety and depression. Its comprehensive nature ensures that patients' health status is evaluated beyond physical symptoms alone. The tool's sensitivity to changes over time renders it suitable for monitoring fluctuations

in CVT patients' conditions.

One other crucial element of the analysis, as shown in Figure 4, is the within-patient correlation. This aspect acknowledges individual variances in baseline quality of life and the diverse trajectories of EQ-5D scores over time. By observing each patient's EQ-5D score progression, the lines' variability underscores the unique paths in quality of life changes among patients, thereby suggesting the inclusion of random effects in the statistical model.

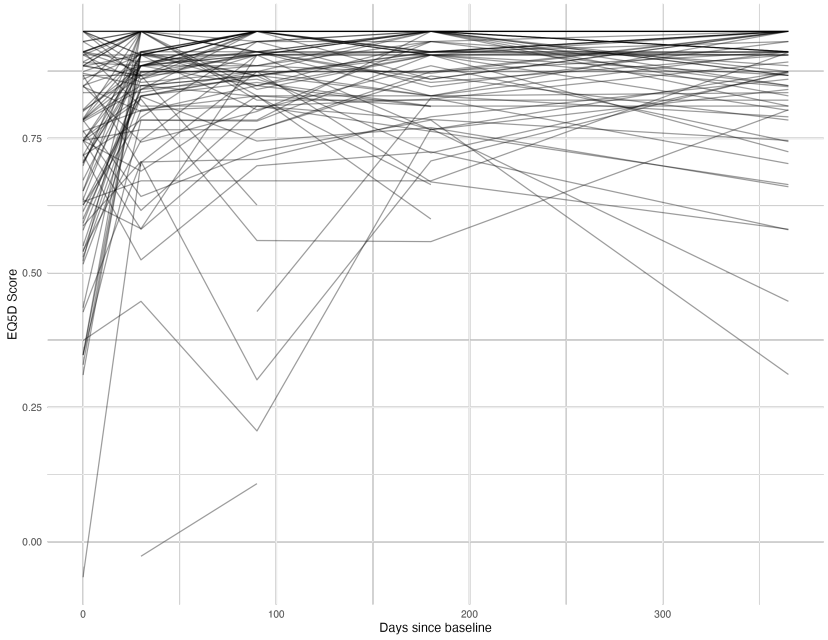

Figure 4: Spaghetti plot of patients' quality of life

An understanding of CVT symptomatology is basic to the extension of the study broad. A graphical depiction of the most prevalent symptoms among participants, as represented in Figure 5, helps identify patterns like the frequent occurrence of headaches compared to the less common neurological symptoms. The bar chart provides a clear comparison of symptom prevalence.

Statistical tests of difference in the proportions of symptoms like 'headache' have shown significant differences, while results for symptoms like 'NV' (nausea/vomiting) revealed no significant difference. Given the absence of strong correlations between pairs of symptom variables for CVT, as indicated by the

phi coefficient (a measure of association between two binary variables) - an investigation into the potential impacts of headaches and 'nausea/vomiting' on CVT outcomes could uncover more details about the phenomenon under study.

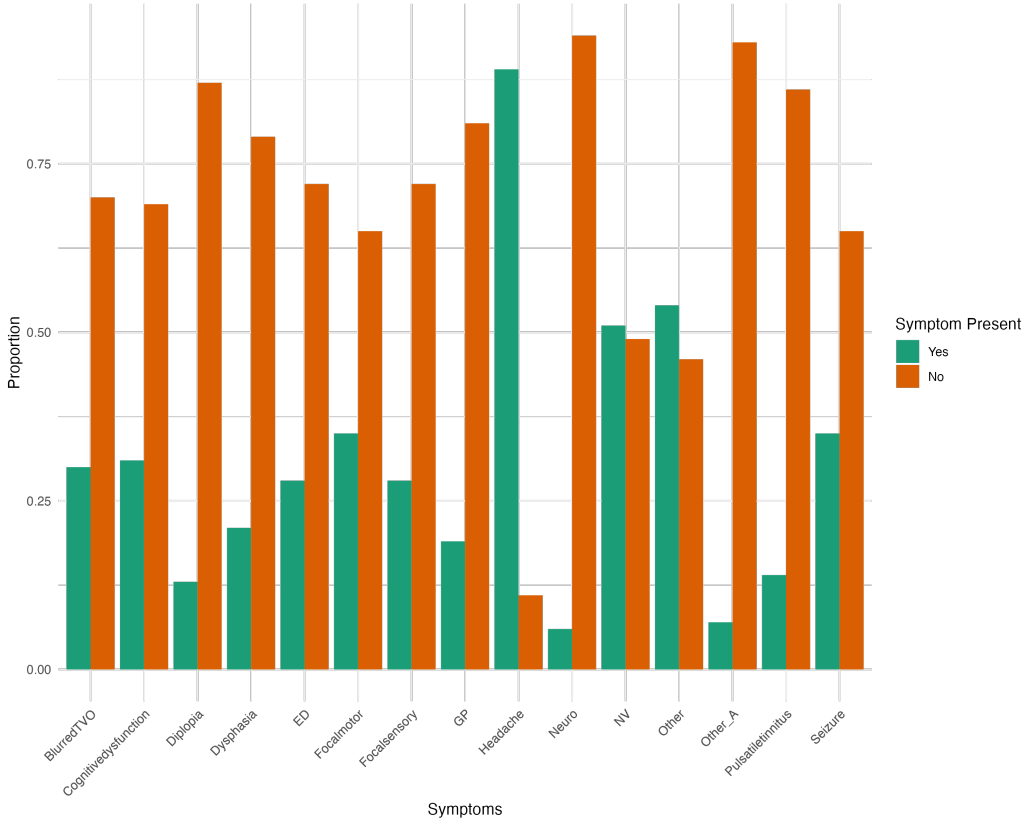

Figure 5: Spaghetti plot of patients' quality of life

5 Formal Analysis

In a comprehensive analysis of cerebral venous thrombosis (CVT) patients, this report portrays the utilization of statistical models to examine the trajectory of patient-reported outcomes (PROMs). Emphasizing the careful investigation into the prevalence of missing data, the study employed logistic regression to evaluate the randomness of missing data, primarily identifying

headache as significantly associated with missing instances. This association provides evidence towards the assumption that the data for the outcome variable EQ-5D may be missing at random (MAR).

The examination of how various factors influence the time until potentially worsening psychological outcomes following a cerebral venous thrombosis (CVT) diagnosis is presented by the Cox proportional hazards model as in (Equation 1). This is especially relevant for the understanding of the impact of delayed diagnosis, which, due to CVT's varied presentation and potential for rapid progression, is an aspect of concern.

The study further offers linear mixed-effects models (LMMs) to facilitate an understanding of individual health trajectories over time, particularly examining how changes in quality of life are influenced by CVT symptoms and educational level. As discussed by [Gabrio et al., 2022], the LMMs effectively address the missing data under MAR without the necessity for imputation, stressing the significance of fixed variables such as headache, other symptoms, nausea/vomiting, and years of education on the EQ5D_Score. The integration of random effects allows for the unique progressions of EQ5D_Score to be captured for each patient, signifying that personalized treatment and improved care are driven by the significant fixed effects that serve as clear clinical targets.

Crucially, the examination compares two distinct models: one considering only random intercepts (Equation 2) and another accounting for both random intercepts and slopes (Equation 3). Discrepancies in the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) as in Table 3A favor the model incorporating random slopes, indicating non-uniform changes in EQ5D_Score over time among the patient cohort.

Alternatively, within the domain of generalized estimating equations (GEE), this report details the exploration of various correlation structures, including 'independence,' 'exchangeable,' and 'unstructured.' The 'independence' structure, by assuming no correlation within patient' repeated measures, is identified as the best model per the Quasi-likelihood under the Independence model Criterion (QIC), which back for a reduction in model complexity when correlation between measures is trivial. The model is in (Equation 4).

6 Results Analysis

By fitting the Cox proportional hazard model, the estimated coefficients are shown in Table 1A. The results indicate that the coefficient for `PHQ9_Score` being negative, suggest higher depressive symptoms are associated with a shorter day to the event of experiencing worse psychological outcome. The hazard ratio ($\exp(\beta_{\text{PHQ9\_Score}}) = 0.97632$) less than 1, implies that as the PHQ9 score increases (indicating more severe symptoms), the hazard, or risk of the event occurring, decreases slightly. This is significant at the 0.05 level ($p = 0.02030$), indicating that the association is unlikely to be due to chance.

Moreover, the positive coefficient for age suggests that older age is associated with a longer day to event. For each year increase in age, the hazard increases by about 1.153% (exp(coef) = 1.01153). This effect is highly significant ($p = 0.00031$), which implies a strong relationship between age and the day to the event.

By emphasis, the negative coefficient for being female indicates a lower hazard ratio compared to males (the reference category). The hazard ratio of 0.58087 suggests that females have about 42% lower risk of the event compared to males. This is highly significant ($p = 2.9 \times 10^{-7}$), suggesting a strong effect of Sex on the day to the event.

For Linear Mixed effect model, by Table 2A the significant, negative coefficients for symptoms 'Other(Yes)' and 'nausea/vomiting(Yes)' suggest that the presence of these symptoms is associated with a lower EQ-5D score, thus a reduced quality of life. With the random effects structure advocating for personalized treatment approaches, the changes in the quality of life, as reported through EQ-5D scores can be attributed to more than just the fixed effects of observed symptoms; they also hinge on unobserved patient-specific factors.

In addition to the symptoms 'Other(Yes)' and 'nausea/vomiting(Yes)' which are significant in LMMs, the total years of education also have a statistical effect on the quality of life with the GEE model Table 3 assuming that the repeated measures on the same patient are uncorrelated.

7 Model Diagnostics

The adequacy of the model can often be visually inspected using residual plots, such as Q-Q plots, which compare observed residual quantiles against theoretically expected quantiles under normality.

Within the scope of this analysis, two models are investigated: the Linear Mixed Model (LMM) and the Generalized Estimating Equations (GEE) model. From Figure 6A the LMM's residual plot indicates a deviation from the expected normal distribution, particularly at the tails, suggesting a potential violation of the normality assumption - a critical consideration in the model's application. On the other hand, the GEE residual plot demonstrates a closer agreement with normality, although with minor deviations, suggesting a better model fit under the assumption of normal residuals.

Given the observed deviations in the LMM residuals, and the lesser deviations in the GEE model, it may be posited that the latter provides a somewhat better fit for this dataset. The inherent fitness of GEE models, particularly their ability to produce consistent estimates without the necessity of normally distributed residuals [McNeish, 2015], further supports for their use in this context.

Furthermore, the GEE model's proficiency for accommodating the correlation structure in resonance to repeated measurements without stringent distributional assumptions on the residuals, highlights its appropriateness for this study's analytical demands.

8 Conclusion

Our model provides an effective tool to identify the outcome trajectories, clinical and neuroradiological factors that may impact these trajectories (PROMs data for baseline, day 30, day 180 and day 365) and targets high-risk individual symptoms for potential decline in the quality of life in patients with cerebral venous thrombosis (CVT).

It also suggests that increasing physical activity may be more beneficial than changing smoking or dietary habits for reducing diabetes risk. model suggests that the number of years invested on education, nausea/vomiting, and some other symptoms which are not very common among individual patients potentially have bad effect on quality of life.

Further studies are needed to confirm the association and to test the

model in different populations and settings.

References

[Gabrio et al., 2022] Gabrio, A., Plumpton, C., Banerjee, S., and Leurent, B. (2022). Linear mixed models to handle missing at random data in trial-based economic evaluations. *Health Economics*, 31(7):1442–1458.

[McNeish, 2015] McNeish, D. (2015). Re: What are the assumptions of the generalized estimating equations? `https://www.researchgate.net/post/What_are_the_assumptions_of_the_generalized_estimating_equations/550710d8d3df3e65508b4613/citation/download`. Retrieved from ResearchGate.

9 Appendix

The Cox proportional hazards model can be mathematically expressed as:

$$h(t) = h_0(t) \exp(\beta_1 \cdot \text{PHQ9\_Score} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Sex}) \tag{1}$$

where:

- $h(t)$ is the hazard at time t,
- $h_0(t)$ is the baseline hazard at time t,
- exp represents the exponential function,
- $\beta_1, \beta_2, \beta_3$ are the coefficients for: the depression severity score {PHQ9 score}, Age, and Sex, respectively.

Table 1A: Time to Event in Cerebral Venous Thrombosis Patients

Variable	**coef**	**exp(coef)**	**se(coef)**	**p-value**
PHQ9_Score	-0.02397	0.97632	0.01033	0.02030 *
Age	0.01147	1.01153	0.00318	0.00031 ***
SexFemale	-0.54322	0.58087	0.10589	2.9e-07 ***

The linear mixed effects model with random effect can be written in the following mathematical form:

$$\begin{aligned} Y_{ij} = \beta_0 &+ \beta_1(\text{Headache}_{ij}) + \beta_2(\text{Other}_{ij}) \\ &+ \beta_3(\text{nausea/vomiting}_{ij}) + \beta_4(\text{TotalYearsOfEducation}_{ij}) \\ &+ b_{0i} + \epsilon_{ij} \end{aligned} \tag{2}$$

where:

- Y_{ij} is the EQ5D_Score for the i-th patient at the j-th observation.
- β_0 is the overall intercept term for the population.

- $\beta_1, \beta_2, \beta_3, \beta_4$ are the fixed effect coefficients for the predictors Headache, Other symptoms, nausea/vomiting, and TotalYearsOfEducation respectively.

- b_{0i} is the random intercept for the i-th patient which accounts for the individual variability in the EQ5D_Score.

- ϵ_{ij} is the residual error term for the i-th patient at the j-th observation.

The notation $\sim 1|\text{ID}$ indicates that the model includes a random intercept for each level of the factor ID, assuming that the b_{0i}'s are normally distributed with mean 0 and some variance σ_b^2.

ϵ_{ij} are assumed to be normally distributed with mean 0 and some variance σ_ϵ^2, and are independent of the b_{0i}. The notation 'na.action = na.exclude' indicates that observations with missing values are excluded from the analysis.

Similarly, the linear mixed-effects model with both random intercepts and slopes can be written as follows:

$$\begin{aligned} Y_{ij} = \beta_0 &+ \beta_1(\text{Headache}_{ij}) + \beta_2(\text{Other}_{ij}) \\ &+ \beta_3(\text{nausea/vomiting}_{ij}) + \beta_4(\text{TotalYearsOfEducation}_{ij}) \\ &+ b_{0i} + b_{1i}(\text{EQ5D}_{ij}) + \epsilon_{ij} \end{aligned} \tag{3}$$

where:

- Y_{ij} is the observed health related quality of life {EQ5D} score for the i-th patient at the j-th day.

- β_0 is the fixed intercept, the average EQ5D score when all covariates are 0.

- $\beta_1, \beta_2, \beta_3, \beta_4$ are the fixed effect coefficients associated with Headache, Other symptoms, NV, and TotalYearsOfEducation, respectively.

- b_{0i} is the random intercept for the i-th patient, accounting for the individual deviation from the average EQ5D score.

- $b_{1i}(\text{EQ5D}_{ij})$ represents the random slope for the i-th patient, which allows the slope (effect of time on EQ5D score) to vary across patients.

- ϵ_{ij} is the residual error term for the i-th patient at the j-th day.

Table 2A: Linear Mixed-Effects Models - Random intercepts and slope

Variables	**Estimate**	**Std. Error**	**p-value**
Intercept	0.997	0.0618	0.0000
Headache(Yes)	-0.024	0.0302	0.4254
Other(Yes)	-0.047	0.0168	0.0066
nausea/vomiting(Yes)	-0.034	0.0171	0.0465
TotalYearsOfEdu	-0.005	0.0034	0.1116

The Generalized Estimating Equations (GEE) model can be expressed mathematically as follows:

$$\begin{aligned}\text{EQ5D\_Score}_{ij} = \beta_0 + \beta_1 \cdot \text{Headache1}_{ij} + \beta_2 \cdot \text{Other1}_{ij} \\ + \beta_3 \cdot \text{nausea/vomiting1}_{ij} + \beta_4 \cdot \text{Totalyearsofeducation}_{ij} + \varepsilon_{ij}\end{aligned} \tag{4}$$

where:

- EQ5D_Score_{ij} is the health-related quality of life score for the i^{th} individual at the j^{th} day.
- β_0 is the intercept, representing the average EQ5D score when all covariates are at zero (assuming the covariates are centered).
- β_1 is the coefficient for the effect of headache.
- β_2 is the coefficient for the effect of other symptoms.
- β_3 is the coefficient for the effect of nausea/vomiting.
- β_4 is the coefficient for the effect of the total years of education.
- ε_{ij} is the error term for the i^{th} individual at the j^{th} day, which is assumed to follow a normal distribution with a mean of zero.

The model uses an "independence" correlation structure, meaning that the repeated measures within individuals are assumed to be uncorrelated.

Table 3: Generalized Estimating Equations (GEE) model

Variable	**Estimate**	**Std. Error**	**p-value**
(Intercept)	0.98632	0.03422	2e-16 ***
Headache (Yes)	-0.03414	0.02223	0.1246
Other (Yes)	-0.04791	0.00980	1e-06 ***
nausea/vomiting (Yes)	-0.03195	0.01146	0.0053 **
Totalyearsofeducation	-0.00522	0.00213	0.0145 *

Table 3A: Comparison of Linear Mixed-Effects Models

Model	**AIC**	**BIC**	**p-value**
(Model 1)	-483	-455	
(Model 2)	-639	-554	0.0001

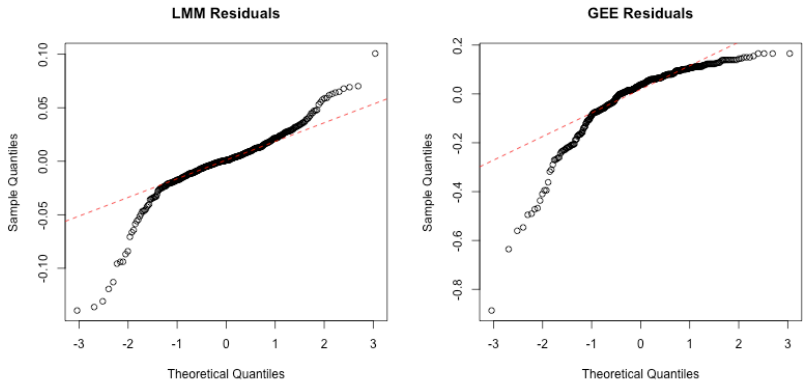

Figure 6A: Q-Q plot for LMM and GEE model

9.1 R Codes

```r
library(ggplot2)
library(tidyverse)
library(reshape2)
library(dplyr)
library(RColorBrewer)

project <- read.csv("project1.csv", header = T)
```

```r

project <- project %>%
  mutate(Sex = factor(Sex)) %>%
  mutate_at(vars(ED:Other), factor)

project <- project[-nrow(project), ] # Removes the last row

# Data wrangling and transformation
# Transforming PHQ9: Depression severity
project_PHQ9 <- melt(project, id.vars = c("ID", "SECRET", "Age", "Gender", "Totalyearsofeducation"),
                     measure.vars = c("PHQ9_BL", "PHQ9_d30", "PHQ9_d90", "PHQ9_d180", "PHQ9_d365"),
                     variable.name = "PHQ9", value.name = "PHQ9_Score")

# Joining additional columns
selected_columns_PHQ9 <- project %>%
  select(ID, Ethnicitycoded:Hemorrhagetype)

project_PHQ9 <- project_PHQ9 %>%
  left_join(selected_columns_PHQ9, by = "ID")

# Transforming EQ-5D-5L: Euro-QoL-5D
project_EQ5D <- melt(project, id.vars = c("ID", "SECRET", "Age", "Sex", "Totalyearsofeducation"),
                     measure.vars = c("EQ5D_BL", "EQ5D_d30", "EQ5D_d90", "EQ5D_d180", "EQ5D_d365"),
                     variable.name = "EQ5D", value.name = "EQ5D_Score")

# Joining additional columns
selected_columns_EQ5D <- project %>%
  select(ID, Ethnicitycoded:Hemorrhagetype)

# Joining additional columns for EQ-5D-5L
project_EQ5D <- project_EQ5D %>%
  left_join(selected_columns_EQ5D, by = "ID")

# Transforming EQ-VAS: Visual Analog Scale
project_EQVAS <- melt(project, id.vars = c("ID", "SECRET", "Age", "Gender", "Totalyearsofeducation"),
                      measure.vars = c("EQVAS_BL", "EQVAS_d30", "EQVAS_d90", "EQVAS_d180", "EQVAS_d365"),
                      variable.name = "EQVAS", value.name = "
```

```
    EQVAS_Score")

# Joining additional columns
selected_columns_EQVAS <- project %>%
  select(ID, Ethnicitycoded:Hemorrhagetype)
# Joining additional columns for EQ-VAS
project_EQVAS <- project_EQVAS %>%
  left_join(selected_columns_EQVAS, by = "ID")

# Transforming HIT-6: Headache Impact Test
project_HIT6 <- melt(project, id.vars = c("ID", "SECRET", "
    Age", "Gender", "Totalyearsofeducation"),
                     measure.vars = c("HIT6_BL", "HIT6_d30",
    "HIT6_d90", "HIT6_d180", "HIT6_d365"),
                     variable.name = "HIT6", value.name = "
    HIT6_Score")

# Joining additional columns
selected_columns_HIT6 <- project %>%
  select(ID, Ethnicitycoded:Hemorrhagetype)
# Joining additional columns for HIT-6
project_HIT6 <- project_HIT6 %>%
  left_join(selected_columns_HIT6, by = "ID")

# Transforming FAS: Fatigue Assessment Score
project_FAS <- melt(project, id.vars = c("ID", "SECRET", "Age
    ", "Gender", "Totalyearsofeducation"),
                    measure.vars = c("FAS_BL", "FAS_d30", "
    FAS_d90", "FAS_d180", "FAS_d365"),
                    variable.name = "FAS", value.name = "FAS_
    Score")
# Joining additional columns
selected_columns_FAS <- project %>%
  select(ID, Ethnicitycoded:Hemorrhagetype)
# Joining additional columns for FAS
project_FAS <- project_FAS %>%
  left_join(selected_columns_FAS, by = "ID")

# Transforming MoCA: Montreal Cognitive Assessment (MoCA)
    scores
project_MoCA <- melt(project, id.vars = c("ID", "SECRET", "
    Age", "Gender", "Totalyearsofeducation"),
                     measure.vars = c("MoCA_BL", "MoCA_d90",
    "MoCA_d180", "MoCA_d365"),
                     variable.name = "MoCA", value.name = "
```

```
   MoCA_score")

# Joining additional columns
selected_columns_MoCA <- project %>%
  select(ID, Ethnicitycoded:Hemorrhagetype)

project_MoCA <- project_MoCA %>%
  left_join(selected_columns_MoCA, by = "ID")

 # Transforming mRS: modified Rankin Scale scores
project_mRS <- melt(project, id.vars = c("ID", "SECRET", "Age
   ", "Gender", "Totalyearsofeducation"),
                    measure.vars = c("mRS_BL", "mRS_d90", "
   mRS_d180", "mRS_d365"),
                    variable.name = "mRS", value.name = "mRS
   _score")

# Assuming you have a project data frame with additional
   columns to join
selected_columns_mRS <- project %>%
  select(ID, Ethnicitycoded:Hemorrhagetype)

# Joining the additional columns with the transformed mRS
   data
project_mRS <- project_mRS %>%
  left_join(selected_columns_mRS, by = "ID")

 # Missing data
# Calculate proportion of missing data for each score in
   their respective datasets
missing_PHQ9 <- sum(is.na(project_PHQ9$PHQ9_Score)) / nrow(
   project_PHQ9)
missing_EQ5D <- sum(is.na(project_EQ5D$EQ5D_Score)) / nrow(
   project_EQ5D)
missing_EQVAS <- sum(is.na(project_EQVAS$EQVAS_Score)) / nrow
   (project_EQVAS)
missing_HIT6 <- sum(is.na(project_HIT6$HIT6_Score)) / nrow(
   project_HIT6)
missing_FAS <- sum(is.na(project_FAS$FAS_Score)) / nrow(
   project_FAS)
missing_MoCA <- sum(is.na(project_MoCA$MoCA_score)) / nrow(
   project_MoCA)
missing_mRS <- sum(is.na(project_mRS$mRS_score)) / nrow(
   project_mRS)

```

```r
# Combine the proportions into a dataframe for plotting

missing_data_df <- data.frame(
  Variable = c("PHQ9_Score", "EQ5D_Score", "EQVAS_Score", "HIT6_Score", "FAS_Score", "MoCA_score", "mRS_score"),
  Proportion = c(missing_PHQ9, missing_EQ5D, missing_EQVAS, missing_HIT6, missing_FAS, missing_MoCA, missing_mRS)
)

# Plot the data
missing_data <- ggplot(missing_data_df, aes(x = Variable, y = Proportion)) +
  geom_segment(aes(x = Variable, xend = Variable, y = 0, yend = Proportion), color = "black") +
  geom_point(aes(x = Variable, y = Proportion), color = "red", size = 3) +
  coord_flip() +
  labs(x = "Outcomes",
       y = "Proportion of Missing Data") +
  theme_minimal()

 # KM Plot - Montreal Cognitive Assessment (MoCA) scores

project_wide <- project_MoCA %>%
  pivot_wider(names_from = MoCA, values_from = MoCA_score)

# Assuming that NA in any of the MoCA scores means the participant dropped out by that time point
# We create a time-to-event data frame.
time_to_event <- project_wide %>%
  mutate(
    stime = case_when(
      is.na(MoCA_d90) ~ 90,
      is.na(MoCA_d180) ~ 180,
      is.na(MoCA_d365) ~ 365,
      TRUE ~ 365  # Censored at the end of the study if no NA found
    ),
    censor = as.integer(!is.na(MoCA_d365))  # 1 if not censored, 0 if censored
  )

# Now we will create the survival object and fit the Kaplan-Meier model.
fit <- survfit(Surv(stime, censor) ~ 1, data = time_to_event)
```

```
# Create the Kaplan-Meier plot using ggfortify's autoplot function
km_plot <- autoplot(fit) +
 labs(x = "\n Time (Days) ", y = "Survival Probability \n") +
 theme(plot.title = element_text(hjust = 0.5),
 axis.title.x = element_text(face="bold", colour="#FF7A33", size = 12),
 axis.title.y = element_text(face="bold", colour="#FF7A33", size = 12),
 legend.title = element_text(face="bold", size = 10))
 # Save the plot
ggsave("km_moca_plot.png", km_plot, width = 10, height = 8, dpi = 300)

# CORRELATION
# Omitting missing values and extracting scores
project_PHQ9_clean <- na.omit(project_PHQ9[c("ID", "PHQ9_Score")])
project_EQ5D_clean <- na.omit(project_EQ5D[c("ID", "EQ5D_Score")])
project_EQVAS_clean <- na.omit(project_EQVAS[c("ID", "EQVAS_Score")])
project_HIT6_clean <- na.omit(project_HIT6[c("ID", "HIT6_Score")])
project_FAS_clean <- na.omit(project_FAS[c("ID", "FAS_Score")])
project_MoCA_clean <- na.omit(project_MoCA[c("ID", "MoCA_score")])
project_mRS_clean <- na.omit(project_mRS[c("ID", "mRS_score")])
# Merging the data frames
project_outcome <- Reduce(function(x, y) merge(x, y, by = "ID", all = TRUE),
                          list(project_PHQ9_clean, project_EQ5D_clean, project_EQVAS_clean, project_HIT6_clean, project_FAS_clean, project_MoCA_clean, project_mRS_clean))


project_outcome <- na.omit(project_outcome)

# Use GGally to create a scatter plot matrix
 ggpairs(project_outcome,
         columns = c("PHQ9_Score", "EQ5D_Score", "EQVAS_Score", "HIT6_Score", "FAS_Score", "mRS_score"),
         upper = list(continuous = wrap("cor", size = 4,
```

```
    method = "pearson")),
177         lower = list(continuous = "smooth"),
178         diag = list(continuous = "barDiag")) +
179   theme_bw() +
180   theme(legend.position = "none") # Remove legend to match
      the uploaded plot style
181
182   # SPAGHETTI PLOT
183 time_mapping <- data.frame(
184   EQ5D = c("EQ5D_BL", "EQ5D_d30", "EQ5D_d90", "EQ5D_d180", "
      EQ5D_d365"),
185   Day = c(0, 30, 90, 180, 365)
186 )
187
188 # Join this mapping to the project_EQ5D to create a 'Day'
      variable
189 project_EQ5D <- project_EQ5D %>%
190   left_join(time_mapping, by = "EQ5D")
191
192 # Convert the EQ5D variable to a factor with levels ordered
      by time to ensure correct plotting
193 project_EQ5D$EQ5D <- factor(project_EQ5D$EQ5D, levels = time_
    mapping$EQ5D)
194
195 # Plotting
196 spaghetti <- ggplot(project_EQ5D, aes(x = Day, y = EQ5D_Score
    , group = ID)) +
197   geom_line(alpha = 0.4) +  # Set alpha for better visibility
      if lines overlap
198   labs(x = "Days since baseline",
199        y = "EQ5D Score",
200        title = "") +
201   theme_minimal() +
202   theme(legend.position = "none")  # Ensuring the legend does
      not appear
203
204 # Barplot for Proportions
205 # The variables of interest
206 vars_of_interest <- c("ED", "Neuro", "GP", "Other_A", "
    Headache", "NV", "BlurredTVO",
207                       "Diplopia", "Focalmotor", "Focalsensory
    ", "Seizure",
208                       "Pulsatiletinnitus", "Dysphasia", "
    Cognitivedysfunction", "Other")
209
```

```r
# Create an empty data frame to store the results
summary_table <- data.frame(Variable = character(),
                            Count_No = integer(),
                            Count_Yes = integer(),
                            Proportion_1 = numeric())

# Loop through each variable and calculate the summary statistics
for (var in vars_of_interest) {
  if (var %in% names(project)) {
    # Count the number of 0's and 1's
    count_0 <- sum(project[[var]] == 0, na.rm = TRUE)
    count_1 <- sum(project[[var]] == 1, na.rm = TRUE)

    # Calculate the proportion of 1's
    proportion_1 <- count_1 / (count_0 + count_1)

    # Add the results to the summary table
    summary_table <- rbind(summary_table, c(Variable = var,
                                            Count_0 = count_0,
                                            Count_1 = count_1,
                                            Proportion_1 = proportion_1))
  }
}

# Convert the summary_table to a data frame
summary_table <- data.frame(summary_table, stringsAsFactors = FALSE)

# Print the summary_table
print(summary_table)
# Create a table for summary statistics with proportions
summary_table <- data.frame(Variable = vars_of_interest,
                            Count_0 = NA_integer_,
                            Count_1 = NA_integer_,
                            Proportion_1 = NA_real_)

for (var in vars_of_interest) {
  summary_table$Count_0[summary_table$Variable == var] <- sum(project[[var]] == 0, na.rm = TRUE)
  summary_table$Count_1[summary_table$Variable == var] <- sum(project[[var]] == 1, na.rm = TRUE)
```

```
  summary_table$Proportion_1[summary_table$Variable == var] <-
    summary_table$Count_1[summary_table$Variable == var] /
    (summary_table$Count_0[summary_table$Variable == var] +
    summary_table$Count_1[summary_table$Variable == var])
}

# Round the Proportion_1 column to 2 decimal places
summary_table$Proportion_1 <- round(summary_table$Proportion_1, 2)

# ARRANGEMENT
# Arrange the summary table in descending order of Proportion_1
summary_table <- summary_table %>%
  arrange(desc(Proportion_1))

# View the ordered summary table
print(summary_table)

summary_table <- summary_table %>%
  mutate(Proportion_2 = 1 - Proportion_1)

# View the summary table with Proportion_2
print(summary_table)

summary_table_long <- melt(summary_table, id.vars = "Variable",
                           measure.vars = c("Proportion_1", "Proportion_2"))

# Define colorblind-friendly colors
cb_friendly_colors <- brewer.pal(n = 2, name = "Dark2")

# Create the bar plot with colorblind-friendly colors
boxplot2 <- ggplot(summary_table_long, aes(x = Variable, y = value, fill = variable)) +
  geom_bar(stat = "identity", position = "dodge") +
  scale_fill_manual(values = cb_friendly_colors,
                    labels = c("Yes", "No"),
                    breaks = c("Proportion_1", "Proportion_2")) +
  labs(x = "Symptoms", y = "Proportion", fill = "Symptom Present") +
  theme_minimal() +
```

```r
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x labels for readability.

  # Test of difference in proportion
# Extract the counts for headache
headache_counts <- summary_table[summary_table$Variable == "Headache", c("Count_0", "Count_1")]

# Run the proportion test for headache
prop_test_result <- prop.test(x = headache_counts$Count_1,
                              n = headache_counts$Count_1 + headache_counts$Count_0)

# Output the result of the proportion test
print(prop_test_result$p.value)

# MISSING AT RANDOM TEST
project_EQ5D$Sex <- factor(project_EQ5D$Sex)
if (length(unique(project_EQ5D$Sex)) < 2) {
  stop("Sex variable has less than two levels.")
}

project_EQ5D <- project_EQ5D %>%
  mutate(Missing_EQ5D_Score = as.numeric(is.na(EQ5D_Score)))

# Fit logistic regression model
mar_model <- glm(Missing_EQ5D_Score ~ Age + Sex + Totalyearsofeducation +
                   Headache + Other + NV,
                 data = project_EQ5D, family = binomial())

# Check model summary
summary(mar_model)

# COX PROPORTIONAL HAZARD
# Fit the Cox proportional hazards model
cox_model <- coxph(Surv(Timefromsymptomstoenrolment) ~ PHQ9_Score + Age + Sex, data = project_PHQ9)

# Summary of the Cox model
summary(cox_model)

# LINEAR MIXED EFFECT MODEL
# Random intercept only

```

```r
fit1_lme <- lme(EQ5D_Score ~ Headache + Other + NV + Totalyearsofeducation,
                random = ~ 1 | ID, data = project_EQ5D, na.action = na.exclude)
summary(fit1_lme)

# Random Intercept and Slope
fit2_lme <- lme(EQ5D_Score ~ Headache + Other + NV + Totalyearsofeducation,
                random = ~ EQ5D | ID, data = project_EQ5D, na.action = na.exclude)
summary(fit2_lme)

# MODEL COMPARISON
anova(fit1_lme,fit2_lme)

par(mfrow = c(1, 2))

# Fit the GEE model
gee_model1 <- geeglm(EQ5D_Score ~ Headache + Other + NV + Totalyearsofeducation,
                     id = ID,
                     data = project_EQ5D_clean,
                     family = gaussian, # Assuming EQ5D_Score is continuous; change if not
                     corstr = "exchangeable") # Choose the appropriate correlation structure

# Summary of the GEE model
summary(gee_model1)

# Fit the GEE model
gee_model2 <- geeglm(EQ5D_Score ~ Headache + Other + NV + Totalyearsofeducation,
                     id = ID,
                     data = project_EQ5D_clean,
                     family = gaussian, # Assuming EQ5D_Score is continuous; change if not
                     corstr = "unstructured") # Choose the appropriate correlation structure

# Summary of the GEE model
summary(gee_model2)

gee_model3 <- geeglm(EQ5D_Score ~ Headache + Other + NV +
```

```
    Totalyearsofeducation,
                      id = ID,
                      data = project_EQ5D_clean,
                      family = gaussian) #  EQ5D_Score is
    continuous
# Summary of the GEE model
summary(gee_model3)

par(mfrow = c(1, 2))

# Define the limits for the plots
lims <- c(-3.5, 3.5)

# QQ plot for LMM residuals
qqnorm(resid(fit2_lme), main = "LMM Residuals")
qqline(resid(fit2_lme), col = "red", lty = 2)

# QQ plot for GEE residuals
qqnorm(resid(gee_model3), main = "GEE Residuals")
qqline(resid(gee_model3), col = "red", lty = 2)

# Reset graphics layout
par(mfrow = c(1, 1))
```