STA 550 Project

Wakeel Kasali

March 8, 2024

Contents

ABSTRACT

Cerebral venous thrombosis (CVT) poses unique challenges for patient management due to the varied clinical manifestations and their impact on quality of life. This study investigates the progression of patient-reported outcomes, particularly depression as measured by the PHQ-9 questionnaire, at critical time points following the CVT event.

A robust exploratory data analysis was conducted to reveal the overall presence and trajectory of key symptoms (headaches, nausea/vomiting) alongside patient demographics and the timing of symptom onset and diagnosis.

Advanced statistical models, including Linear Mixed-Effects Models (LMMs) and Generalized Estimating Equations (GEEs), were employed to quantify the influence of these factors on depression severity over time. These analyses not only highlighted the individual variations in patient experiences but also emphasized the need for a personalized approach to treatment.

The findings of this study offer crucial insights for optimizing clinical strategies and improving the quality of life for CVT patients. This work represents a significant step towards achieving tailored patient care.

1 Introduction

Cerebral venous thrombosis (CVT) is a condition primarily affecting young women, often linked to birth control use 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 measures to track symptom changes over a year.

In light of this, some of the statistical questions that would be answered include:

- Is there a statistically significant association between the presence of initial symptoms (like headache, nausea/vomiting, blurred vision, diplopia) and the severity of depression (PHQ9 scores) over the follow-up period?
- How do clinical features such as venous infarct, midline shift, and hemorrhage at diagnosis predict the trajectory of PHQ9 scores across the different time points (baseline, day 30, day 180, and day 365)?

- Are there age-related differences in the impact of CVT symptoms on the quality of life and psychological well-being as measured by PHQ9 scores?

This research will improve our understanding of CVT's long-term effects and pave the way for better patient care strategies.

2 Data description and summaries

The data collection occurred across a span of three years, featuring a complete 12-month follow-up of individual participant. The process incorporated consistent assessments through the use of patient-reported outcome measures (PROMs) and neuroimaging techniques. In total, the study involved 103 patients with 53 participating in the clinical trial and another 50 registered in the registry.

This dataset comprises a comprehensive variety of variables arranged as follows: Demographic and clinical variables include: ID, which is a unique identifier (Categorical); Age (Continuous); Sex/Gender (Categorical); Ethnicitycoded, (Categorical); Dateofenrolment, (Date format); Timefromsymptomstoenrolment and Symptomstodiagnosis, which are time-related variables measured in days (Continuous).

The section on Symptoms and Clinical Assessments covers variables representing the presence or absence of symptoms like headache, nausea/vomiting (NV), blurred vision, diplopia, etc. (Categorical - Binary). It also includes NIHSS (Continuous); findings from neuroimaging such as Venousinfarct, Midlineshift, and Hemorrhage (Categorical); and mRS Scores, referring to the Modified Rankin Scale scores, used for measuring disability (Continuous).

The study includes (PROMs) which encompass scores for assessing headache, mood, fatigue, cognition, and quality of life across various time points (Continuous). This also includes EQ5D and EQVAS, which are EuroQol measures for health-related quality of life (Continuous), as well as FAS, HIT6, and PHQ9, which are specific scales for fatigue, headache impact, and depression severity (Continuous). Treatment variables include Anticoagulant, which refers to the type of anticoagulant medication used (Categorical), and Recanalization Status, assessed at different time intervals (Categorical).

Missing values are present in the dataset, particularly evident in cognitive data compared to other outcomes.

3 Exploratory analysis

A robust analysis hinges on complete data. Visualization (like in Figure 1A) can be instrumental in pinpointing missing value proportions within key study variables. To understand the CVT study better, a table summarizing patient symptoms and diagnoses, similar to Table 1, is recommended. This table quantifies symptom prevalence and frequency, offering valuable insights into their impact on patient well-being.

Table 1: Summary of Categorical Variables

Part 1				Part 2			
Variable	Count_0	Count_1	Prop_1	Variable	Count_0	Count_1	Prop_1
ED	74	29	0.28	Seizure	67	36	0.35
Neuro	97	6	0.06	Puls_tinn	89	14	0.14
GP	83	20	0.19	Dysphasia	81	22	0.21
Oth_A	96	7	0.07	Cog_dysfn	71	32	0.31
Headache	11	92	0.89	Other	47	56	0.54
NV	50	53	0.51				
Blrd_TVO	72	31	0.30				
Diplopia	90	13	0.13				
Foc_motor	67	36	0.35				
Foc_sensory	74	29	0.28				

For CVT patients, particularly those experiencing headaches (the most common symptom), a boxplot representation of PHQ-9 scores (like Figure 2A) is a powerful tool. This visualization allows for assessing changes in depression severity over time and gauge treatment effectiveness.

Independent of group analysis, tracking individual depression score changes over time within the SECRET and TOP-SECRET groups is facilitated by a line graph visualization (as seen in Figure 1). This approach aids in identifying potential differences in symptom trajectory or treatment outcomes between the groups.

Furthermore, a correlation heatmap is crucial for discerning relationships between symptom factors so also between some demographic factors. This can reveal important influences on the recovery process, informing further analysis. Figures 3A and 2, analyzing both continuous and categorical variables, unveil non-strong pairwise correlations. This suggests that individual

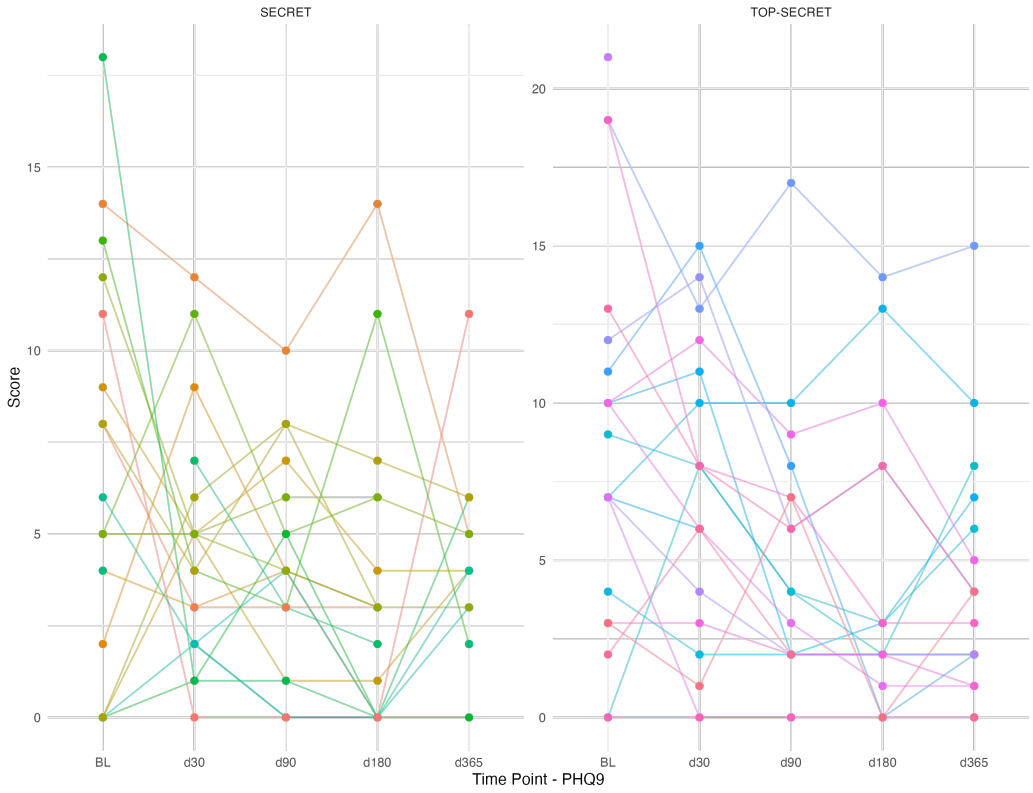

Figure 1: Trend Lines for PHQ-9 over Time for 20 Sampled IDs

variables, rather than strong interactions between them, may each have a unique influence on patient outcomes.

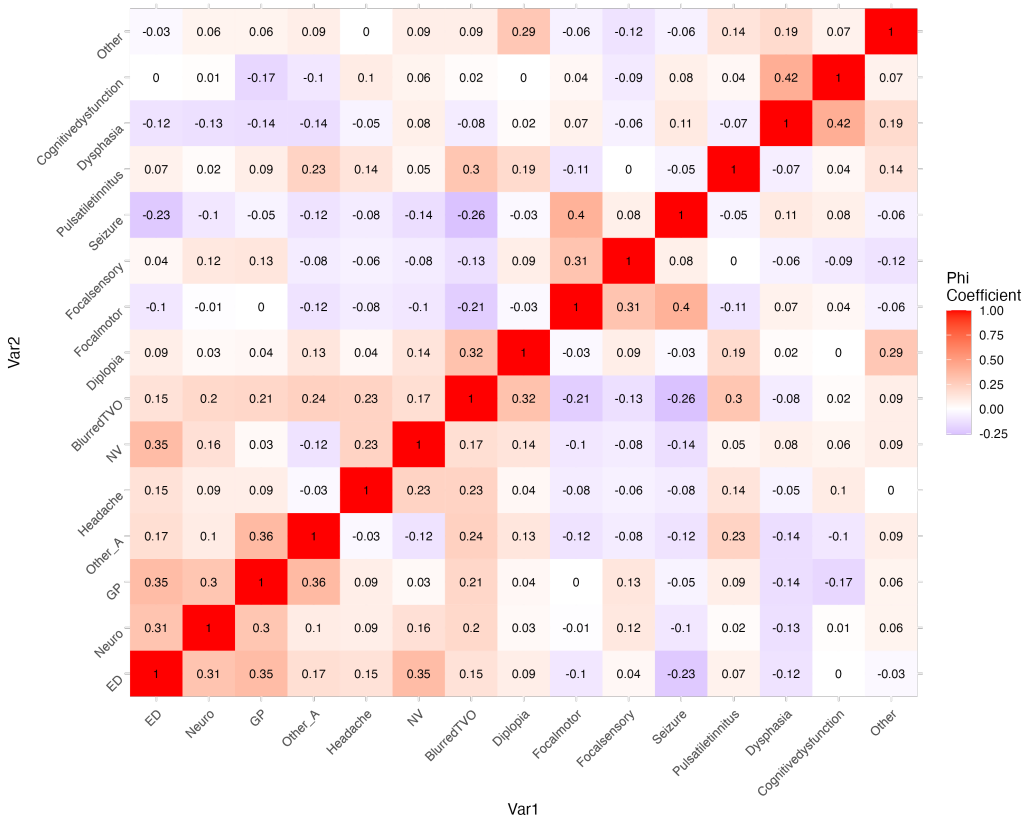

Figure 2: Correlation heatmap - Categorical variables

4 Formal Analysis

Two statistical methods are recommended to investigate the relationship between clinical/neuroradiological data and patient-reported symptoms of depression in CVT patients. These methods are Linear Mixed-Effects Models (LMMs) and Generalized Estimating Equations (GEEs).

LMMs (as in appendix equation 1) offer an advantage over linear regression by incorporating both fixed effects, which are consistent across patients, and random effects, which account for individual variations. This approach is particularly well-suited for analyzing longitudinal data, such as the repeated PHQ-9 depression scores collected in this study.

GEEs (as in appendix equation 2) are particularly adept at handling data with inherent correlations. These correlations arise from repeatedly

measuring the same variable within a single patient. GEEs provide flexibility by accommodating various correlation structures within the data, even when the exact structure is unknown. This flexibility, combined with robust standard errors, ensures reliable results when investigating population-level trends and the effects of various factors on patient outcomes.

The high prevalence of headaches and nausea/vomiting (NV) among CVT patients, along with available data on patient demographics and symptom timing, presents an opportunity to infer the trajectory of depression severity (reflected by PHQ-9 scores) over time.

By a way forward, a two-step LMM approach is recommended. In the first step, a model would be fitted with only a random intercept, accounting for baseline differences in depression scores between patients. The second step would involve fitting a model incorporating both a random intercept and a random slope. By comparing these two models, we can assess whether the inclusion of a random slope, which represents individual differences in the rate of change over time, significantly improves the model's fit to the data.

Linear Mixed-Effects Models (LMMs) are employed in this study to elucidate the average influence of various factors on depression severity, as measured by the PHQ-9. These factors encompass the prevalence of symptoms like headaches (β_1) and nausea/vomiting (NV) (β_2), alongside patient characteristics such as age (β_3) and education level. The analysis focuses on fixed effects associated with each variable (β coefficients), which quantify the average change in PHQ-9 scores across all patients. For instance, the fixed effect for headaches (β_1) will reveal the average increase or decrease in depression score linked to experiencing headaches.

However, to account for inherent variability in how individual patients experience depression, LMMs incorporate a random intercept (u_{0j}). This term captures patient-specific deviations from the overall population average. In simpler terms, some patients may have inherently higher or lower baseline depression scores, even before considering the effects of the other variables. This random intercept component allows for a more nuanced understanding of how these factors influence depression severity in CVT patients.

Similar to LMMs, Generalized Estimating Equations (GEEs) are employed to reveal the direction and strength of the relationships between symptoms, demographic and clinical variables, and the depression scores (PHQ-9). However, the interpretation of the coefficients differs between these two methods.

In the context of GEEs, the coefficients are interpreted as representing the average change in depression scores across the entire CVT patient population for a one-unit change in the predictor variable. This approach focuses on population-level trends, rather than individual-level changes captured by the fixed effects in LMMs.

5 Discussion

While composite indices, such as the modified Rankin Scale, Euro-QoL-5D, EQ-VAS: Visual Analog Scale, Headache Impact Test, and Fatigue Assessment Score, offer a broad perspective, important changes within specific domains relevant to this study, particularly mood, might be obscured by their use. The PHQ-9 specifically addresses this limitation by providing a focused assessment of depression. This makes the PHQ-9 (Population health questionnaire (depression)) a suitable choice to be employed as the outcome variable.

A robust foundation for understanding CVT patient outcomes was established through exploratory data analysis. Visualizations and summary tables revealed the prevalence of key symptoms (headaches, nausea/vomiting) and the variability in depression severity (PHQ-9 scores). Analyses employing boxplots and line graphs highlighted distinct patterns and trajectories of symptom progression, emphasizing the need for personalized treatment approaches.

Formal analyses utilizing Linear Mixed-Effects Models (LMMs) and Generalized Estimating Equations (GEEs) further solidified these observations. LMMs quantified the average influence of symptoms and patient characteristics on depression scores, while accounting for individual variability through the incorporation of random intercepts. GEEs complemented this by revealing population-level trends in these relationships.

Critically, insights from both exploratory and formal analyses converged on the multifaceted nature of CVT outcomes. Non-strong correlations between variables were identified, suggesting that individual factors independently impact recovery and quality of life. This finding underscores the need for a multi-dimensional approach to patient care, tailored to the diverse and evolving needs of CVT survivors.

By employing sophisticated statistical models, this study lays the groundwork for the development of more nuanced clinical strategies and future re-

search endeavors. The ultimate goal remains to improve the well-being of patients affected by this challenging condition.

6 Appendix

$$
\begin{aligned}
y_{ij} =& \beta_0 + \beta_1 \times \text{Headache}_{ij} + \beta_2 \times \text{NV}_{ij} \\
&+ \beta_3 \times \text{Age}_{ij} + \beta_4 \times \text{TotalyearsOfEducation}_{ij} \\
&+ \beta_5 \times \text{TimefromSymptomsToEnrolment}_{ij} + \beta_6 \times \text{SymptomsToDiagnosis}_{ij} \\
&+ u_{0j} + \epsilon_{ij}
\end{aligned} \tag{1}
$$

The β coefficients represent the fixed effects of factors like the presence of headaches (Headache), NV, age, education level, symptom duration from onset to enrollment, and time to diagnosis.

The term u_{0j} represents the random intercept for each patient (j), accounting for individual variations in baseline depression scores.

Finally, ϵ_{ij} represents the random error associated with each individual measurement.

$$
E(y_i) = g(\mathbf{z}_i^T \boldsymbol{\beta}), \quad \text{Cov}(y_i) = \Sigma_i, \tag{2}
$$

where $g(x) = h^{-1}(x)$ and $g(\mathbf{z}_i^T \boldsymbol{\beta}) = \mathbf{z}_i^T \boldsymbol{\beta}$ for linear models.

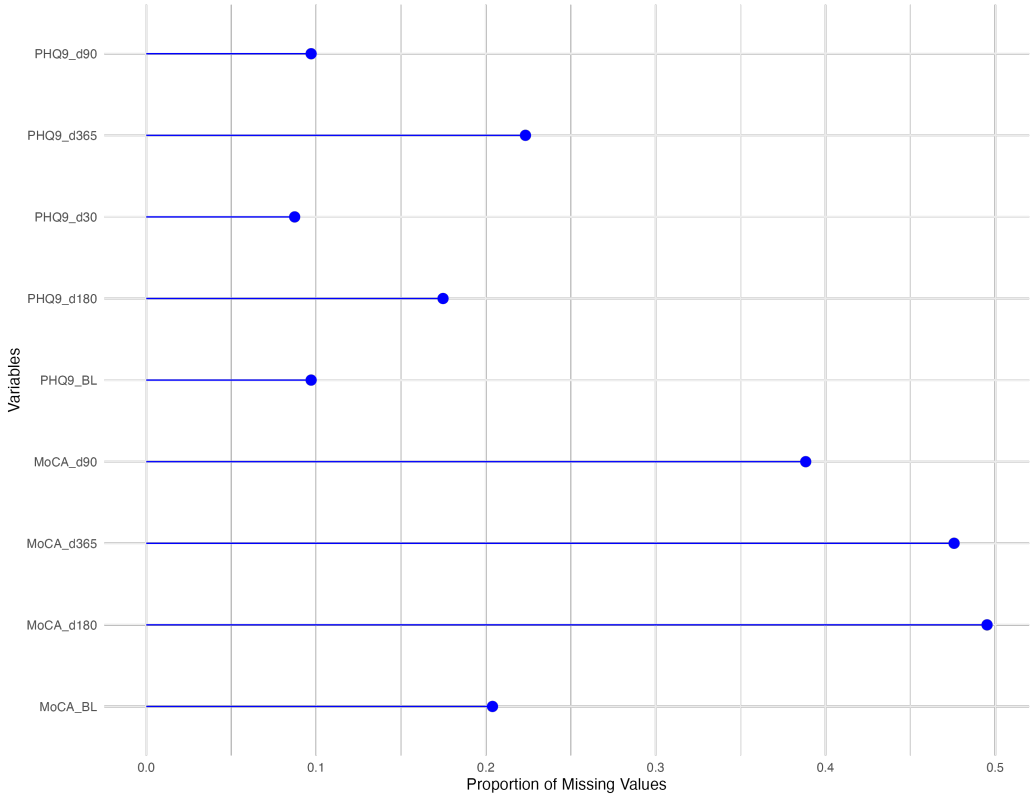

Figure 1A: PHQ9 Scores for Patients with Headache at Different Time Points

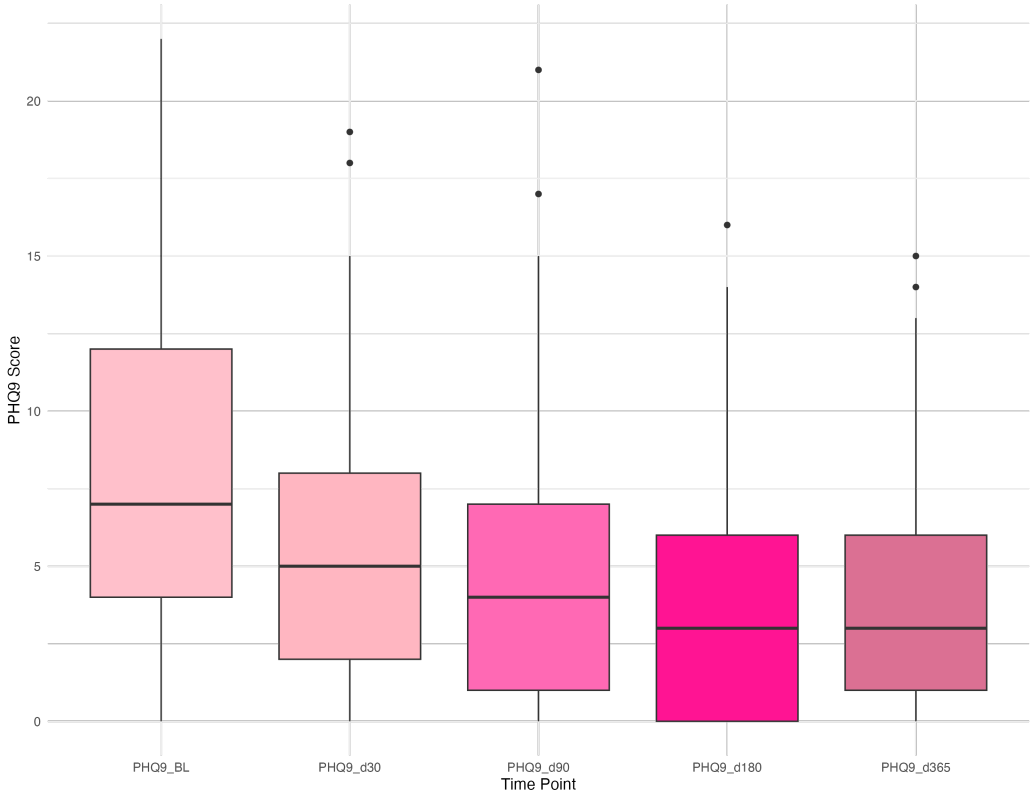

Figure 2A: Patients PHQ-9 trajectories for headache symptom

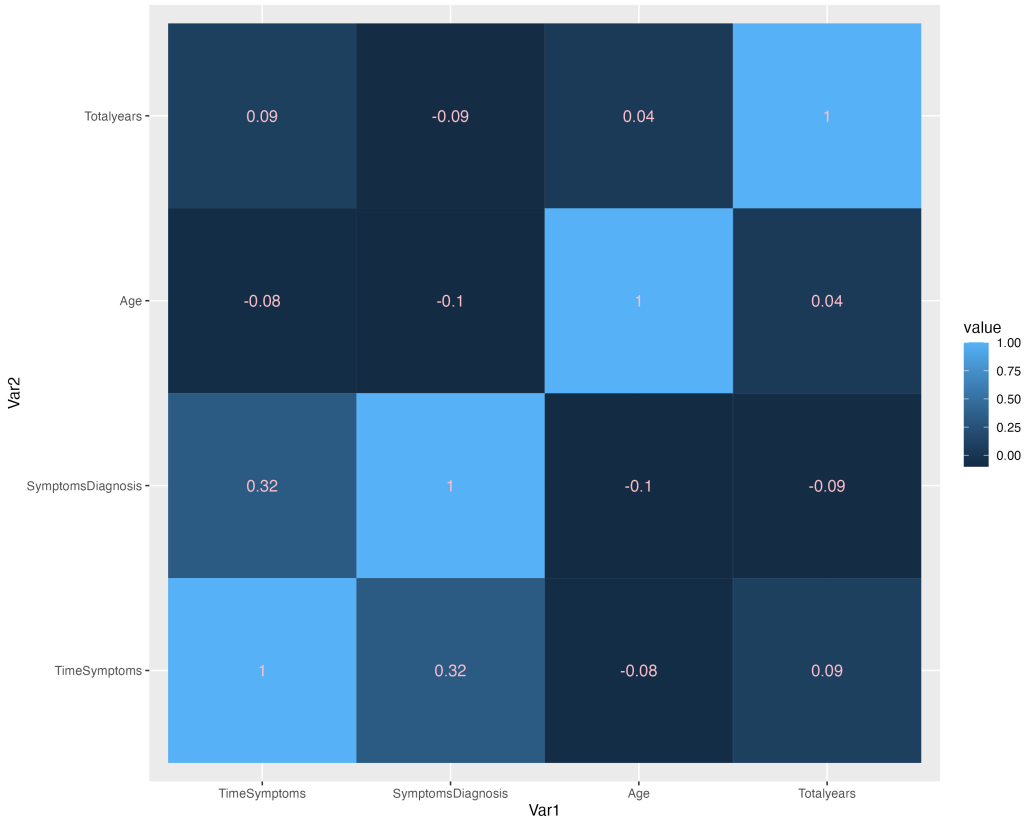

Figure 3A: Correlation heatmap - Continuous variables