Analysis of Tasks 0 & 1: Descriptive Analysis and Data Preprocessing

Effectiveness of Anti-CGRP Monoclonal Antibodies in Migraine Treatment

Course: Artificial Intelligence & Healthcare Statistics

December 18, 2025

1 Introduction

This document outlines the rationale behind the R code implementation for Tasks 0 and 1. It bridges the gap between raw data manipulation and clinical relevance, preparing the dataset for rigorous statistical modeling (Mixed Models and Survival Analysis).

2 Data Loading and Inspection

Code Chunk: Sections 1, 2, 3 (Load Data, Dimensions, Structure, Summary)

2.1 Technical Rationale

Before conducting any analysis, data integrity must be verified. This involves inspecting the dataset dimensions ($N \times P$) to ensure no truncation occurred during the import process and verifying variable types (e.g., distinguishing between integers and character strings).

2.2 Clinical Relevance

The summary statistics generated in this phase serve as the initial "sanity check" of the study population.

- **Range Verification:** Verifying that variables such as `AGE` fall within expected biological ranges (e.g., mean ≈ 51.3 years) ensures the exclusion of corrupted entries.
- **Real-world Data Nature:** This step highlights the inherent noise in observational data, where outliers and inconsistencies are more prevalent compared to controlled clinical trials (RCTs).

3 Missing Data Pattern Visualization (Task 1)

Code Chunk: Section 3b (`naniar`, `gg_miss_var`, `md.pattern`)

3.1 Technical Rationale

This section addresses the exploration of missingness patterns. Determining the mechanism of missing data is crucial for selecting the correct handling strategy:

- **MCAR:** Missing Completely at Random.
- **MAR:** Missing at Random.

- **MNAR:** Missing Not at Random.

3.2 Clinical Relevance

In longitudinal migraine studies, missingness is rarely random.

- **Pattern Check:** Frequent missingness in Monthly Migraine Days (MMDs) during later cycles (e.g., 3rd cycle) often suggests patient dropout due to lack of efficacy or adverse events.
- **Imputation Strategy:** Visualizing missingness in covariates (e.g., `BMI`, `Sleep Disorders`) informs the decision between listwise deletion (dropping rows) and advanced imputation methods like MICE to preserve statistical power.

4 Preprocessing: Type Conversion

Code Chunk: Section 4 (`mutate`, `as.factor`)

4.1 Technical Rationale

R defaults to reading numerical codes in CSV files (e.g., Sex = 1, 2) as continuous integers. This is mathematically incorrect for regression models, which would interpret `"2"` as having twice the value of `"1"`. Converting these variables to **Factors** creates dummy variables (0/1 indicators) appropriate for model fitting.

4.2 Clinical Relevance

Correct labeling is vital for interpretation:

- **Antibodies:** Distinguishing between *Erenumab*, *Galcanezumab*, and *Fremanezumab* allows for comparative efficacy testing in Task 2.
- **Diagnosis:** Separating *Chronic Migraine* from *Medication Overuse Headache (MOH)* is essential, as MOH patients typically exhibit more resistant disease trajectories.

5 Data Merging (Master Dataset)

Code Chunk: Section 5 (`left_join`)

5.1 Technical Rationale

A `left_join` operation is performed to merge **longitudinal** (time-varying) data with **baseline** (static) data, using `SUBJECT_ID` as the key.

5.2 Clinical Relevance

This structure is a prerequisite for Longitudinal Analysis. To determine which baseline characteristics predict response (Q2), the Linear Mixed-Effects Models (Task 2) require access to static baseline covariates (like `AGE` and `BMI`) for every monthly `MMD` entry.

6 Descriptive Analysis: Table 1 (Task 0)

Code Chunk: Section 6 (`table1`)

6.1 Technical Rationale

This generates the standard "Table 1" found in clinical research, stratifying data by treatment group (`ANTIBODY`) to assess randomization balance.

6.2 Clinical Relevance

This fulfills the inspection of baseline distributions to identify potential confounders.

> *Example:* If the Erenumab group has a significantly higher baseline BMI or prevalence of Sleep Disorders, differences in treatment outcomes might be attributable to these comorbidities rather than the drug itself.

7 Longitudinal Trajectory Plot (Task 0 & Q1)

Code Chunk: Section 7 (`ggplot`, `geom_line`)

7.1 Technical Rationale

This plots the mean MMDs over time, including Standard Error (SE) bars to visualize variability.

7.2 Clinical Relevance

This addresses Q1: *Do baseline MMDs progressively decrease over time?*

- **Visualizing Efficacy:** Expected patterns include a "sawtooth" or continuous decline. A flattening curve after Cycle 1 ($C1_{end}$) indicates a plateau effect.
- **Wearing-off Effect:** Coloring by Cycle helps identify phenomena where efficacy wanes toward the end of a treatment administration period.

8 Dropout Analysis (Task 0)

Code Chunk: Section 8 (`table`, `prop.table`)

8.1 Technical Rationale

Calculates the frequency of the `Suspension` variable.

8.2 Clinical Relevance

Quantifying dropouts is critical for defining the study population:

- **Validity:** High dropout rates ($> 30\%$) can invalidate results.
- **Causality:** Distinguishing between adverse events vs. lack of efficacy is necessary.
- **Populations:** Helps define "Intent-to-Treat" (ITT) vs. "Per-Protocol" (PP) populations.

9 Imputation Comparison (Task 1)

Code Chunk: Section 9 (LOCF vs. Linear Interpolation)

9.1 Technical Rationale

Comparisons are made between two single-imputation methods for time-series data:

1. **LOCF (Last Observation Carried Forward):** Assumes the patient's condition remains stable from the last observed time point.
2. **Linear Interpolation:** Draws a straight line between two known points.

9.2 Clinical Relevance

Given that migraine is a fluctuating condition:

- **Critique of LOCF:** While often "conservative" in progressive diseases, it may bias results here. If a patient drops out due to worsening headaches, LOCF captures the high frequency.
- **Critique of Interpolation:** If data is missing due to random events (e.g., vacation), linear interpolation may better approximate natural MMD fluctuation.

10 Baseline Imputation with MICE (Task 1)

Code Chunk: Section 9b (`mice`)

10.1 Technical Rationale

This utilizes **Multivariate Imputation by Chained Equations (MICE)**. Unlike simple mean imputation, MICE models each missing value conditionally based on other variables in the dataset.

10.2 Clinical Relevance

This is essential for Task 2 (Predictors). Simple deletion of rows with missing comorbidities could result in a 20-30% loss of the cohort. MICE creates complete datasets that preserve the correlation structure (e.g., the relationship between Obesity and Migraine severity).

11 Responder Rate Calculation (Q1)

Code Chunk: Section 10 (`Pct_Reduction`, `Responder_50`)

11.1 Technical Rationale

The outcome variable is derived using the following formula:

$$\text{Pct Reduction} = \frac{\text{MMD}_{\text{baseline}} - \text{MMD}_{\text{current}}}{\text{MMD}_{\text{baseline}}} \times 100 \tag{1}$$

11.2 Clinical Relevance

This addresses Q1 regarding the proportion of patients achieving $\geq 30\%$ and $\geq 50\%$ reduction.

- **Regulatory Standard:** The $\geq 50\%$ responder rate is the gold standard endpoint for migraine trials.
- **Clinical Significance:** While a raw reduction of 2 days might be statistically significant, it may not be clinically meaningful. A 50% reduction represents a tangible improvement in quality of life. The 30% threshold is often reserved for "difficult-to-treat" chronic patients.

12 Key Takeaways

- **Foundation Established:** Data cleaning, type conversion, and demographic summaries (Table 1) are complete.
- **Missing Data Strategy:** Missingness patterns have been visualized, and MICE has been selected as the robust imputation method for regression tasks.
- **Outcome Defined:** The calculation of the $\geq 50\%$ responder rate prepares the dataset for Survival Analysis (Time to 50% response).