Analysis of Task 3: Temporal Data Mining and Electronic Phenotyping

February 5, 2026

Abstract

This section addresses Research Question 3 (Q3), applying unsupervised machine learning to identify distinct patient phenotypes from longitudinal data. Using K-Means clustering on temporally abstracted features, we identified three clinically relevant subgroups: "Standard Responders," "Super Responders," and "Refractory Patients." These phenotypes provide a granular stratification method beyond simple baseline characteristics.

1 Introduction

While Linear Mixed-Effects Models (Task 2) estimate the *average* population trajectory, they may mask heterogeneity within the cohort. The objective of Task 3 is to discover whether distinct patient phenotypes emerge from the data and if these computational phenotypes can serve as tools for patient stratification [1].

2 Methodology: Temporal Abstraction & Clustering

2.1 Feature Engineering

To convert complex time-series data into clusterable features, we applied **temporal abstraction**. Rather than using raw monthly values, we engineered high-level features for each patient to capture the dynamics of their disease burden:

- **Level:** Mean Monthly Migraine Days (MMDs) per Cycle (C1, C2, C3).
- **Volatility:** Standard deviation of MMDs, measuring the stability versus fluctuation of symptoms.
- **Slope:** The linear rate of improvement over the 3-year period.

2.2 Determining Optimal Clusters

We employed the Elbow Method to determine the optimal number of phenotypes (k).

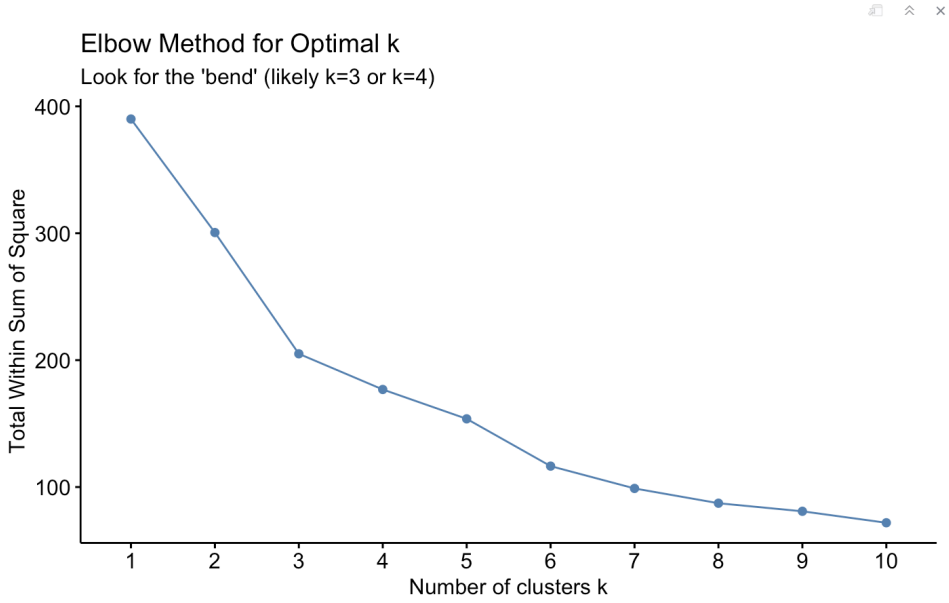
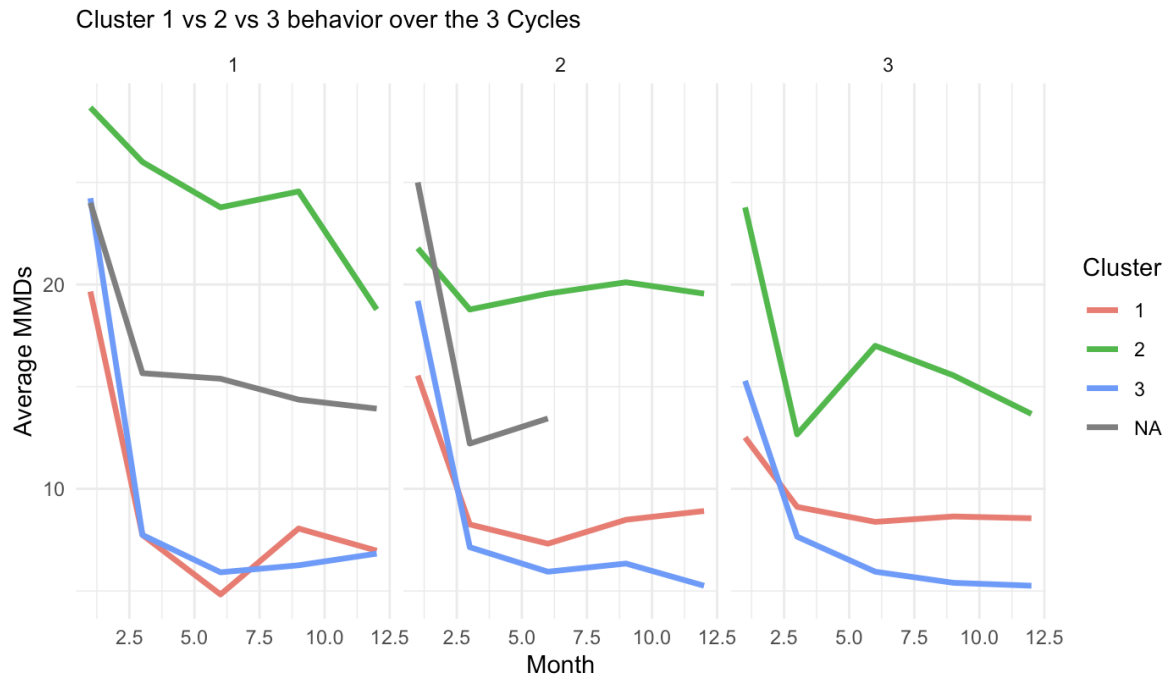


Figure 1: Elbow Method for Optimal k. The inflection point at $k = 3$ indicates the optimal balance between model complexity and interpretability.

As shown in Figure 1, the "Total Within Sum of Square" drops sharply from $k = 1$ to $k = 3$ and begins to plateau thereafter. Consequently, we selected $k = 3$ as the optimal number of clusters.

3 Results: Identification of Migraine Phenotypes

Applying K-Means clustering ($k = 3$) revealed three distinct patient profiles. These are visualized in the longitudinal trajectory plot (Figure 2).




```
;. Clinical Interpretation (Validation)
```

Figure 2: Migraine Trajectories by Computational Phenotype over 3 Cycles.

3.1 Phenotype Characterization

Based on the visual patterns, we characterized the clusters as follows:

1. **Cluster 1 (Red): "Standard Responders" (Moderate Baseline)**
   - *Trajectory:* Patients start with a moderate burden (≈ 20 days). They exhibit a rapid response in Cycle 1, dropping to < 10 days, and maintain this low level throughout the study.
   - *Clinical Profile:* This represents the "ideal" stable responder.
2. **Cluster 2 (Green): "Refractory Phenotype" (High Burden)**
   - *Trajectory:* Patients start with the highest burden (≈ 28 days). Despite some initial reduction, MMDs remain consistently high (> 15 days). They exhibit severe "rebound" spikes at the start of each cycle.
   - *Clinical Profile:* These are Non-Responders or "Difficult-to-Treat" patients who fail to achieve low frequency despite continuous therapy.
3. **Cluster 3 (Blue): "Super Responders" (Deep Remission)**
   - *Trajectory:* Patients start with a high burden (≈ 25 days), similar to the Refractory group. However, they experience a dramatic and sustained drop, reaching the lowest levels of the entire cohort (< 5 days).
   - *Clinical Profile:* This group demonstrates "Deep Remission," proving that high baseline severity does not preclude excellent outcomes.

4 Clinical Stratification (Validation)

We assessed whether baseline characteristics could predict these cluster assignments.

4.1 Stratification by Severity and Age

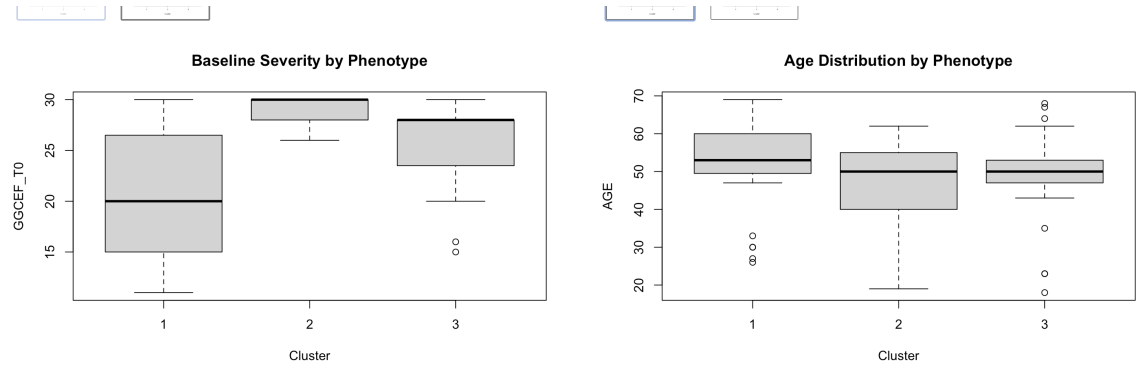


Figure 3: Distribution of Baseline Severity (Left) and Age (Right) across Phenotypes.

The analysis reveals a **"Severity Paradox"**:

- **Standard Responders (Cluster 1)** have distinctively lower median baseline severity (≈ 20).
- **Refractory (Cluster 2) and Super Responders (Cluster 3)** share similarly high median baseline severity ($\approx 28 - 30$).

Interpretation: Baseline severity is a necessary but insufficient predictor. While low severity predicts Cluster 1, high severity is ambiguous—it may lead to treatment failure (Refractory) or deep remission (Super Responder). Age distribution was uniform across clusters, offering no discriminatory power.

5 Key Takeaways

1. **Distinct Trajectories:** The cohort is not uniform; it splits into three distinct phenotypes (Standard, Refractory, Super Responder).
2. **Clinical Utility:** Identifying Cluster 2 (Refractory) early is critical, as these patients incur high treatment burden with minimal benefit.
3. **Prediction Limit:** Phenotypic success cannot be predicted by disease severity alone, suggesting unmeasured biomarkers distinguish the Super Responders from the Refractory group.

References

[1] Project Schema 2025-26: Medical Application and Healthcare Module.