

Dataset Summary & Initial Diagnostics

Linglu Li

Attribute Definitions

The cross-sectional MRI dataset used in this study contains 436 adults aged 18–96, including both men and women. Among these, 100 subjects over age 60 were clinically diagnosed with very mild to moderate Alzheimer’s disease, and a subset of 20 nondemented participants were rescanned within 90 days for reliability assessment. The outcome variable Group2 was derived from the CDR score, with $CDR = 0$ coded as ‘Nondemented’ and all other CDR values coded as ‘Demented’.

Table 1: Dataset description.

| No. | Attributes | Description |
|-----|------------|---|
| 1 | ID | Identification |
| 2 | M/F | Gender (M if Male, F if Female) |
| 3 | Hand | Handedness |
| 4 | Age | Age in years |
| 5 | EDUC | Years of education |
| 6 | SES | Socio Economic Status (Score from 1-5) |
| 7 | MMSE | Mini Mental State Examination (Score from 0-30) |
| 8 | CDR | Clinical Dementia Rating (Scale from 0-1) |
| 9 | eTIV | Estimated Total Intracranial Volume |
| 10 | nWBV | Normalize Whole Brain Volume |
| 11 | ASF | Atlas Scaling Factor |
| 12 | Group2 | Dementia status (Nondemented vs. Demented) |
| 13 | Delay | Delay |

Data Structure

The cross-sectional dataset contained substantial missingness in several variables, including EDUC (201 values; $\approx 46\%$), MMSE (201 values; $\approx 46\%$), CDR (201 values; $\approx 46\%$), and SES (220 values; $\approx 50\%$) were missings. Because a valid dementia label cannot be assigned without CDR, all cases with missing CDR were removed prior to analysis. These same rows also lacked EDUC and MMSE, so removing them eliminated missingness in those variables as well. After this filtering step, SES was the only variable with remaining missing values (19 cases), which were imputed using KNN with $k = 5$. The full imputation procedure and associated R code are detailed in the Methods section.

Table 2: Variable types and structure of the dataset.

| Variable | Type | Missing | Unique_Values |
|----------|-----------|---------|---------------|
| ID | character | 0 | 436 |
| M.F | character | 0 | 2 |
| Hand | character | 0 | 1 |
| Age | integer | 0 | 73 |
| Educ | integer | 201 | 6 |
| SES | integer | 220 | 6 |
| MMSE | integer | 201 | 18 |
| CDR | numeric | 201 | 5 |
| eTIV | integer | 0 | 312 |
| nWBV | numeric | 0 | 182 |
| ASF | numeric | 0 | 282 |
| Delay | character | 0 | 15 |

Descriptive Statistics

Table 3 provides descriptive statistics for the primary clinical and imaging variables, including their minimum, maximum, mean, and median values. Figure 1 provides boxplots of the primary continuous attributes, displaying their quartile ranges (Q1–Q3), median, mean, and potential outliers. This graphical view helps reveal distributional characteristics such as skewness and variability that are not fully captured by summary statistics alone.

Table 3: TABLE 3 | Min, max, and median values of each attribute.

| Variable | Min | Max | Mean | Median |
|----------|----------|----------|--------------|----------|
| Educ | 1.000 | 5.000 | 3.1787234 | 3.000 |
| SES | 1.000 | 5.000 | 2.5191489 | 2.000 |
| MMSE | 14.000 | 30.000 | 27.0638298 | 29.000 |
| CDR | 0.000 | 2.000 | 0.2851064 | 0.000 |
| eTIV | 1123.000 | 1992.000 | 1459.4978723 | 1447.000 |
| nWBV | 0.644 | 0.847 | 0.7491319 | 0.747 |
| ASF | 0.881 | 1.563 | 1.2165106 | 1.213 |

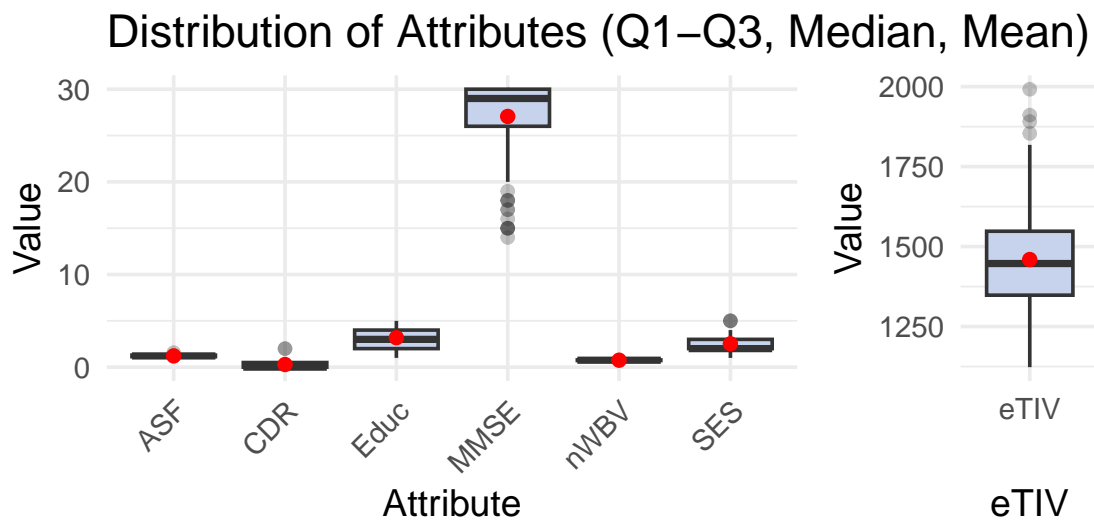


Figure 1: Boxplot distribution of continuous attributes.