Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Uncovering Latent Cognitive Subgroups via MMSE Scores in a Bangladeshi Dementia Cohort: An integration of Data Coresets and Ranked Set Sampling in Gaussian Mixture Modeling

Sharmin Akther

sharminakther12.ju@gmail.com

Jahangirnagar University

Azizur Rahman

Government of Manitoba

Rumana Rois

Jahangirnagar University

Article

Keywords: Alzheimer's disease, Gaussian Mixture Models, Mini-Mental State Examination, Cognitive subgroups

Posted Date: March 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8996904/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Background

Alzheimer's disease (AD) has an incidence of 57 million people worldwide, with it being expected to increase further in low and middle-income nations (LMICs), where Bangladesh is reported to have a prevalence of 8.0% of dementia among persons aged ≥ 60 years and older. Although widely used in cognitive screening, the Mini-Mental State Examination (MMSE) exhibits variable sensitivity (23–76% for mild cognitive impairment) and significant age-education effects, which preclude the use of fixed cutoff strategies to define cognitive subgroups. New evidence suggests that AD has heterogeneous subtypes with different cognitive pathways, other than homogeneous disease progression.

Methods

This article used Gaussian Mixture Models (GMMs) to determine latent cognitive subgroups among a Bangladeshi AD cohort (N = 663) of the National Institute of Neurosciences and Hospital, Dhaka (January 2019-August 2024). The scores with MMSE were divided into severe dementia (MMSE 0–10), moderate dementia (MMSE 11–20), mild cognitive impairment (MMSE 21–25), and normal cognition (MMSE > 25). Two computationally efficient subsampling methods were compared: Ranked Set Sampling (RSS; N = 320) and coreset construction (N = 198), on the overlap minimization and distributional fidelity. Four-component GMMs were estimated using the Expectation-Maximization algorithm on the entire dataset and the two subsamples. The fitted models were evaluated based on log-likelihood values, convergence behavior ($\varepsilon = 10^{-8}$), and pairwise overlap percentages between components.

Results

The MMSE Score had significant pair-wise overlap between the neighboring severity elements (46.5% severe-moderate) as an expression of linear cognitive decline. RSS (N = 320) focused on mild impairment (58.28%), with a low degree of overlap (1.2%), but with a serious lack of severe cases (7.32%). The balanced severity representation (severe: 11.61% of original data; 30%, which contains coreset sampling) and moderate overlaps (18.23%) and weighted mean (19.79) were closest to the population mean (19.74).

Conclusions

Coreset sampling was better than the Ranked Set Sampling because it maintained the severity representation balance in all the levels of cognitive impairment with only 30% of the initial data. The

scalable method facilitates effective CD (cognitive disease) subtype identification in resource-limited environments, supporting enhanced clinical trial design and individualized risk assessment.

1. Introduction

Alzheimer's disease (AD) is the most common form of dementia, which is estimated to affect 57 million individuals around the world and is expected to rise to 153 million by 2050, although the highest rises are found in low- and middle-income countries (LMICs) (1, 2). A cross-sectional study in Bangladesh, a fast aging south Asian country of more than 170 million individuals, has reported a prevalence of dementia at 8.0% of those aged 60 years of age or above, with exceptional differences by age, sex, education, and geographical location (3). Poor levels of literacy and limited availability of neurologists also increase the difficulties in diagnosing in the area, which is why there is a necessity to use a screening instrument that is easy to access and has a high psychometric quality (4, 5).

One of the most popular instruments that is used in the resource-constrained environment is the Mini-Mental State Examination (MMSE), a 30-point assessment instrument of cognitive screening, because of its simplicity and rapidity (6). The meta-analytic evidence suggests that the MMSE has a pooled sensitivity of 79.8% and specificity of 81.3% in identifying dementia in memory clinic practices, but has a sensitivity of only 23–76% in revealing mild cognitive impairment (MCI) (7, 8). Such variable diagnostics, in combination with well-reported age and education effects on MMSE performance, emphasize the weakness of using fixed cutoffs to determine cognitive severity (9).

There is an increasing body of literature to suggest that AD does not have one uniform pattern of progression, but instead represents several separate subtypes with varied cognitive development, biomarker, and reaction to treatment (10, 11). Gaussian Mixture Models (GMMs) are becoming an important set of probabilistic tools that can determine latent cognitive subgroups in heterogeneous populations. In contrast with hard-assignment clustering algorithms, GMMs not only estimate posterior membership probabilities of overlapping components, but also model the continuous range of cognitive decline naturally modeled (12). Models using GMMs and other latent-class models of longitudinal AD data have effectively determined different classes of trajectories with differing cognitive decline rates and risk of MCI-to-dementia conversion that differ by odds ratios of 0.12–3.02 across the classes found (13, 14).

Despite these developments, the overlap of GMM elements remains a primary limitation and is especially acute in LMIC cohorts, where the study sample underrepresents severe cases of impairment due to participation bias and a lack of disease duration in severe cases (15, 16). This asymmetry undermines maximum likelihood estimation and biases the estimates of parameters, and decreases the sensitivity to minor subpopulations (17). Strategic subsampling offers a means to reduce computational burden and component overlap through methods such as coreset construction (18–20) and ranked set sampling (21, 22), though effectiveness depends critically on the sampling mechanism employed,

making the choice of sampling strategy of paramount importance for preserving distributional fidelity in mixture modeling applications (23, 24).

Ranked Set Sampling (RSS), a method that was originally applied in agricultural estimation, uses auxiliary rankings to be more representative than simple random sampling, but concentrates on middle distributional ranges and under-represents tails (25, 26). Coreset algorithms are based on computational geometry and build provably representative weighted subsets, whose size is independent of the underlying data, and which are shown to be better than uniform sampling to preserve cluster structure and tail representation (27, 28). These two subsampling strategies of GMM-based subgroup identification have never been directly compared in a previous study investigating an LMIC AD cohort.

This paper fills this research gap with Bangladeshi AD cohort data. We have trained four-component GMMs using the entire dataset, and subsamples created using RSS and coreset. We aimed to: (1) measure component overlap and proportion shifts across different strategies relative to one another; (2) evaluate the distortion of clinical MMSE severity means and tail representation; and (3) identify which subsampling method maximizes the reduction in overlap with distributional and clinical fidelity. The results will inform scalable resource-efficient strategies for the stratification of AD subtypes in LMICs.

This paper is organized into four major sections. Section 2 details the materials and methods, i.e., the data preparation, ethical approval, and our method of analysis, i.e., Gaussian mixture modeling, ranked set sampling, and coreset construction. Section 3 contains the results, which depict the patterns of MMSE distribution, model-based cognitive subgroups, and correlations between different methods of sampling. Section 4 offers a discussion of the clinical and methodological implications, which makes sense of subgroup continuity and sampling effectiveness. Section 5 sums up the main findings and gives recommendations on further directions.

2. Methods and Materials

2.1. Study Design and Data Source

In this study, cross-sectional data on Alzheimer's disease (AD) in the Bangladeshi population have been used, and the Mini-Mental State Examination (MMSE) scores were used as the variable of interest. The dataset included 687 AD patients, who were tested within the period of January 2019 and August 2024 at the National Institute of Neurosciences and Hospital (NINS&H), Dhaka, Bangladesh. Informed consent was signed, and the essential ethical approval was received from the Institutional Review Board of NINS&H, with an approval number of (IRB no. NINS/2023/312). The inclusion criteria included participants aged 50 years or older, diagnosed with AD using the DSM-5 criteria, whereas the exclusion criteria involved participants with comorbidities in the field of neurology, psychiatric disorders, or missing MMSE data. In addition to MMSE scores, the following variables were collected: demographic characteristics (age, sex, district, residence, educational status, occupation, marital status, economic status), clinical symptoms (memory loss, depression, slow gait, anxiety), medical history (family history

of dementia, hypertension, diabetes mellitus, chronic kidney disease, Parkinson's disease, multiple system atrophy),

2.2. Data Preparation and Cleaning

There were originally 687 observations of patients with Alzheimer's disease in the raw data. After stringent preprocessing in line with our previous publication,(29) all the missing scores in the Mini-Mental State Examination (MMSE) scores and any invalid scores (scores beyond the valid range of 0–30) were dropped. Upon this cleaning, the final analytical sample was 663 participants whose MMSE had full and valid data. These MMSE scores were then further placed in four groups of clinical severity with common cutoffs (30, 31): severe dementia (MMSE 0–10), moderate dementia (MMSE 11–20), mild cognitive impairment (MMSE 21–25), and normal cognition (MMSE > 25). This preprocessing stage guaranteed data integrity, fullness, and preparedness to further model development and the subsampling-based estimation process. The overall study design is summarized in Fig. 1, which illustrates the analytical pipeline from data preparation through Gaussian mixture modeling using both Ranked Set Sampling and coreset approaches.

2.3. Randomization and Reproducibility

A fixed random seed was used to ensure complete reproducibility of all statistical procedures, and when all sampling and modeling processes were started. This seed was also used throughout Ranked Set Sampling (RSS) initialization, coreset construction, and Gaussian Mixture Model (GMM) parameter initialization with the Expectation-Maximization (EM) algorithm, thus making all results completely reproducible.

2.4. Sampling Strategy

2.4.1 Ranked Set Sampling (RSS)

Ranked Set Sampling (RSS) is a proven sampling method of ordered data that we used to effectively estimate the underlying distribution of MMSE scores at the cost of a smaller computational burden (32, 33). Similar to McIntyre (1952) (33) and Chen et al. (2004) (32), the entire sample (N = 663) was utilized in creating 80 independent sampling cycles. Each cycle was conducted in four sets (m = 4), each set comprised of four individuals, and these sets were randomly formed in a manner that their scores were ranked based on age, which served as an auxiliary variable due to its correlation with MMSE scores (r = -0.20), whereby higher age rankings corresponded to greater cognitive impairment. Only a single person (reflecting a predetermined rank position) was picked out of a set (e.g., the top-ranked member of the first set, the second-ranked member of the second set, etc.). This process ended up with the final sample size of 320 measured individuals (80 cycles × 4 sets), or 48.5% of the original data.

2.4.2 Coreset Construction

More to the point, to achieve better computational efficiency of the mixture modeling, we have applied a coreset-based approximation in the style of Feldman et al. (34). Coresets are weighted subsets that maintain the statistical characteristics of the entire data with conceptual guarantees, especially on large-scale mixture modelling (35, 36). In order to build a representative coreset and at the same time retain the clinical structure of the MMSE data, we used a category-weighted stratified sample strategy using MMSE clinical categories. Sampling probability of each category was proportionate to the prevalence of the category in the entire dataset, which preserved the underlying distribution of cognitive states. Each sampled observation was assigned a weight $w_i = \frac{N_c}{m_c \cdotN}$, where N_c denotes the number of individuals in category *c*, m_c is the number of sampled points from that category, *N = 663* is the total dataset size, and m is the coreset size. The coreset size was determined using,

$$m = \min\left(\max\left(50, \lceil k \cdot \frac{\log N}{\epsilon^2} \rceil\right), \lfloor 0.3N \rfloor\right)$$

1

with *k* = 4 representing the target number of components and *ε* = 0.1 denoting the approximation error tolerance, in accordance with the theoretical bounds proposed by Bachem et al.(37).

2.5. Statistical Modeling Approach

2.5.1 Gaussian Mixture Modeling

To identify latent subgroups within the MMSE distribution, we applied finite Gaussian Mixture Models (GMM) using the Expectation-Maximization (EM) algorithm (38). The probability density function for a k-component mixture is:

$$f\left(x|\,\theta\right) = \sum_{i=1}^{k} \lambda_i \varphi(x|\mu_i, \sigma_i^2)$$

2

Where, λ_i are mixing proportions ($\sum \lambda_i = 1$), $\varphi(\cdot|\mu_i, \sigma_i^2)$ is the normal density with mean μ_i and variance $\sigma_i^2, \theta = \{(\,\lambda_i, \mu_i, \sigma_i^2), i = 1, \ldots, k\}$ are model parameters

2.5.2 Weighted EM for Coresets

For coreset-based estimation, we employed weighted EM updates as implemented in the **mixtools** package (39):

$$\mu_i^{(t+1)} = \frac{\sum_{j=1}^{m} w_j \tau_{ij}^{(t)} x_j}{\sum_{j=1}^{m} w_j \tau_{ij}^{(t)}}$$

$$\sigma_i^{2(t+1)} = \frac{\sum_{j=1}^{m} w_j \tau_{ij}^{(t)} (x_j - \mu_i^{(t+1)})^2}{\sum_{j=1}^{m} w_j \tau_{ij}^{(t)}}$$

3

$$\lambda_i^{(t+1)} = \frac{\sum_{j=1}^{m} w_j \tau_{ij}^{(t)}}{\sum_{j=1}^{m} w_j}$$

where w_j are coreset weights and $\tau_{ij}^{(t)}$ are posterior probabilities.

2.5.3 Component Overlap Analysis

To quantify the separation between mixture components, we calculated pairwise overlap areas using:

$$\text{Overlap}_{ij} = \int_{-\infty}^{\infty} \min\left[\lambda_i \varphi(\, x|\mu_i, \sigma_i^2), \lambda_j \varphi\,(x|\,\mu_j, \sigma_j^2\,)\right] dx$$

4

This metric ranges from 0 (complete separation) to 1 (complete overlap), providing quantitative evidence of distributional continuity versus discrete categorization.

2.6. Validation Framework

2.6.1 Coreset Validation

The coreset approximation was tested through the assessment of the capability to maintain the structural properties of the original MMSE distribution. The coreset subsample was fitted to the four-component Gaussian Mixture Model through the Expectation-Maximization algorithm, and the component parameters, such as the mixing proportions, means, and standard deviations, were obtained. Pairwise overlap integrals were used to measure component overlap to determine separation between latent subgroups. To test how the multimodal structure was preserved, visual inspection of the fitted density curves and the component distributions has been conducted. These studies offered a descriptive evaluation of the ability of the coreset to capture the underlying distributional characteristics of the entire data, with specific focus on the ability to represent the entire cognitive range of severe impairment to almost normal cognition.

2.6.2 RSS Validation

The Ranked Set Sampling (RSS) procedure was evaluated using several criteria. First, rank balance

was assessed by examining the distribution of selected rank positions across all sampling cycles, with each rank (1 through 4) expected to appear approximately 80 times in the final sample of 320 observations. Second, ranking quality was evaluated by comparing the mean MMSE scores across the

four rank positions; since age was used as the auxiliary ranking variable with a negative correlation to MMSE (r = -0.20), rank 1 (youngest) was expected to have the highest mean MMSE, while rank 4 (oldest) was expected to have the lowest mean MMSE. Third, distributional comparability was assessed by comparing summary statistics (mean, standard deviation, minimum, maximum) of the RSS sample with those of the original dataset, and by fitting a four-component Gaussian Mixture Model to the RSS subsample and comparing the resulting component parameters (proportions, means, standard deviations) with those obtained from the full dataset. Finally, sampling stability was ensured by fixing the random seed (2026) at the beginning of the RSS procedure, guaranteeing reproducibility of all sampling results.

2.6.3 Model Robustness

To ensure the robustness and reliability of the proposed modeling framework, several validation strategies were employed. The EM algorithm was executed with a maximum of 1,000 iterations and a convergence threshold of $\varepsilon = 10^{-8}$. The same random seed (2026) was always used in all analyses- RSS sampling, coreset construction, and GMM initialization- to ensure maximum reproducibility of all the results. The stability of the parameters was measured based on the estimated mixture components (proportions, means, and standard deviations) of the three sampling strategies so that they could be afforded an opportunity to estimate the extent to which each of the sampling strategies retained the distributional structure of the original cognitive data. Also, component overlap was measured using the pair-wise overlap integrals so as to determine the distance between latent subgroups and the continuousness of cognitive decline among different levels of severity. These group controls helped to guarantee that the modeling model gave dependable and readable results, even with the smaller sample sizes of the subsampling methods. R software (version 4.4.1) was used to do all the analyses.

2.7. Ethical Considerations

The study plan was formally approved by the Institutional Review Board (IRB) of the National Institute of Neurosciences and Hospital (Approval No: NINS/2023/312). All of the study participants were informed by way of written consent. In the case of subjects who had serious cognitive impairment, informed consent was given by legally authorized representatives, and further consent by the subjects wherever feasible.

3. Results

3.1. Descriptive Analysis

Based on our data of 687 participants, we deleted 24 participants whose scores on the Mini-Mental State Examination (MMSE) were not present, leaving us with an analytical sample of 663 participants (losses: 3.49%). This complete case data was then used for all subsequent analyses. The four categories of cognitive impairment based on the MMSE scores were severe (0–10), moderate (11–20), mild (21–25), and normal (26 and higher). Table 1 summarizes the distribution and descriptive statistics

of the MMSE scores between these categories. Most of the respondents belonged to moderate impairment of the cognitive status (42.8%), and mild impairment (28.2%). The mean scores of the MMSE showed a declining pattern with the severity of the disorders, and the severe and moderate categories showed variation. This increased variability likely reflects the heterogeneous nature of cognitive decline in advanced disease stages, where patients may present with differing rates of progression, varying domain-specific impairments, and the influence of comorbid conditions, resulting in a wider range of MMSE scores within these severity groups.

Table 1
Descriptive statistics of MMSE scores across cognitive impairment categories (N = 663).

MMSE Category	N	Proportion (%)	Mean	SD	SE	Median	Q1	Q3	IQR	Min	Max
Severe	57	8.6	6.67	3.1	0.41	8	5	9	4	0	10
Moderate	284	42.8	16.4	2.7	0.16	17	14	19	5	11	20
Mild	187	28.2	22.9	1.5	0.11	23	22	24	2	21	25
Normal	135	20.4	27.9	1.4	0.12	28	27	29	2	26	30

In Fig. 2, the distribution of MMSE scores is shown among the four categories of cognitive impairment: Severe (MMSE 0–10), Moderate (MMSE 11–20), Mild (MMSE 21–25), and Normal (MMSE ≥ 26). The severe group has the lowest median score with moderate variability, and the moderate group has a larger interquartile range of scores, which depicts that there is more dispersion in the scores. On the contrary, the mild and normal groups are also tighter around medians (IQR = 2 for both), which indicates that within-group variability is lower. The visual plot supports the progressive loss of the cognitive scores, with the severity of impairment, which are summarized in Table 1.

3.1.2 Normality Assessment of MMSE Scores

Figure 3 displays the normality test results for Mini-Mental State Examination (MMSE) scores in the study population. The left panel represents the kernel density plot of the observed MMSE scores (blue bars) superimposed on the theoretical normal density curve (red line) based on the sample mean and standard deviation. The distribution also exhibits multimodality, negative skewness of the left, and a pronounced peak in the range of moderate impairment, which is significantly different from the overlaying normal distribution. The null hypothesis of normality was rejected by the Shapiro-Wilk test ($W = 0.97005$, $p = 2.14 \times 10^{-10}$). The normal Q-Q plot is given on the right panel, where the quantiles of the sample are compared to the quantiles that should appear according to a normal distribution. There are clear deviations from the line of reference, especially in both tails and in the central part (in the form of an S-shaped curve), which proves the high level of non-normality and the existence of non-homogeneous subpopulations, which are the representatives of various levels of cognitive impairment. These findings suggest that the MMSE scores of this cohort of Alzheimer patients are not distributed normally, and it is

appropriate to use non-parametric statistics or finite mixture models to conduct additional statistical analysis.

3.2. Gaussian Mixture Model Fit to the Full Dataset (N = 663)

Approximate parameter estimates were obtained using a four-component Gaussian Mixture Model (GMM) with Mini-Mental State Examination (MMSE) scores as the data (N = 663). Model selection was guided by comparing fit indices across 1 to 6 components. The 2-component model yielded a log-likelihood of -2154.5 (AIC = 4318.9, BIC = 4341.4), while the 3-component model showed improved fit (log-likelihood = -2138.1, AIC = 4292.1, BIC = 4328.1). The 4-component model achieved a log-likelihood of -2137.7 with AIC = 4297.5 and BIC = 4346.9, offering a balance between model fit and parsimony. Although the 5-component model produced a slightly higher log-likelihood (-2120.4), it exhibited convergence issues with repeated warnings of variances approaching zero, indicating overfitting and unstable component estimation. Models with 1 and 6 components failed to converge altogether. Thus, the 4-component solution was selected as the optimal model, offering the best trade-off among statistical fit, numerical stability, and clinical interpretability, with each component corresponding to a meaningful level of cognitive impairment, ranging from severe to near-normal functioning.

The expectation-maximization algorithm was used to fit the model, whose components were ranked in terms of mean to approximate the clinical progression of severity of the condition (severe to normal cognitive functioning). Each of the components is defined by its approximate mixing ratio (percent), mean MMSE score, standard deviation (SD), and minimum and maximum values.

The MMSE distribution showed a multimodal pattern, indicating that the study population comprises multiple distinct cognitive subgroups rather than a single homogeneous group. The MMSE distribution was multimodal, which was well-fit by the fitted model. The highest proportion (28.2%) was in Component 1, which had a mean of 13.34 (SD = 5.8), which is within the moderate range of impairment, but with a few lower scores, as there is more heterogeneity among this group. The largest subpopulation with a significant overlap with the other levels of severity indicated that component 2 (49.5%) had a mean of 20.21 (SD = 3.9), which represents the upper range of *moderate* and lower range of mild impairment. The small, highly concentrated cluster with a mean of 21.48 (SD = 0.6) referred to component 3 (2.4%), representing a small part of the range of *mild* impairments. Component 4 (19.9%) generated a mean of 27.41 (SD = 1.9), which is near the *normal* cognitive range with a comparatively low standard deviation.

These estimates show fair overall consistency with the empirical category means of Table 1, as well as showing the inherent continuity and partial overlap of stages of cognitive impairment as modeled by the GMM. The entire-dataset fit parameters are the reference (gold-standard) parameters with which the performance of subsampled models (Ranked set and Coreset) is compared in the following parts.

Table 2
Component-wise Distributional Characteristics Estimated from the Gaussian Mixture Model Fit to the Full Dataset (N = 663).

Component	Proportion (%)	Mean	SD	Minimum	Maximum
Component 1	28.2	13.34	5.8	2.1	24.6
Component 2	49.5	20.21	3.9	12.5	27.9
Component 3	2.4	21.48	0.6	20.3	22.7
Component 4	19.9	27.41	1.9	23.7	31.1

The histogram of mini-mental state examination (MMSE) scores is shown in Fig. 4 with the fitted four-component Gaussian mixture model overlaid (the black solid line is the overall mixture density line). Densities of individual components are presented in different colors: Component 1 (green), Component 2 (orange), Component 3 (red), and Component 4 (purple). Vertical dashed lines represent the means of the components estimated. Colored areas of the background are associated with four categories of cognitive impairment that are clinically determined (Severe: 0–10, Moderate: 11–20, Mild: 21–25, Normal: 26–30). The model is an excellent representation of the multimodal distribution of MMSE scores, with its components agreeing fairly with the severity groups of clinical meaning, but with significant interdependence between adjacent groups, especially in the moderate-to-mild spectrum.

3.3. Performance of Gaussian Mixture Models Under Different Sampling Strategies

3.3.1. Ranked Set Sampling (RSS)

Table 3 summarizes the distributional properties of the four-component Gaussian Mixture Model applied to the Ranked Set Sampling (RSS) subset of MMSE data (N = 320, obtained from 80 cycles with 4 sets each containing 4 individuals), where age was used as the auxiliary ranking variable due to its correlation with MMSE score (r = -0.20, available for all 660 participants). We also constructed a composite variable, Dementia Risk Score, derived from Diagnosis and memory loss indicators; however, its correlation with MMSE score was relatively weak (r = -0.15) and, importantly, was only available for 556 participants due to missing data, further supporting the selection of age as a completer and more reliable auxiliary variable for ranking in the RSS procedure.

The RSS sample is dominated by Component 3 (58.28%), representing moderate impairment with a mean MMSE of 19.44 (SD = 3.8), followed by Component 1 (23.91%) with a mean of 10.89 (SD = 5.7), capturing the severe-to-moderate range. Component 2 (10.49%) corresponds to mild impairment (mean = 26.15, SD = 1.61), while Component 4 (7.32%) represents near-normal cognition (mean = 29.64, SD = 0.5). These findings demonstrate that RSS, with age as the ranking variable, effectively concentrated sampling effort on the central portion of the cognitive spectrum (Components 1 and 3), which together comprise over 82% of the sample. However, the procedure underrepresents both extremes, particularly

the near-normal group (Component 4) and, to a lesser extent, the severely impaired range, highlighting the inherent bias of RSS toward the middle of the distribution when the auxiliary variable (age) is only moderately correlated with the outcome of interest (MMSE score). This trade-off underscores both the efficiency and the limitations of ranked set sampling in this Alzheimer's disease cohort.

Table 3
Distributional Characteristics of Gaussian Mixture Model Components Obtained via Ranked Set Sampling (N = 320).

Component	Proportion (%)	Mean	SD	Minimum	Maximum
Component 1	23.91	10.89	5.7	0	13
Component 2	10.49	26.15	1.61	26	28
Component 3	58.28	19.44	3.81	14	25
Component 4	7.32	29.64	0.5	29	30

Figure 5 displays the fitted four-component Gaussian Mixture Model (GMM) density overlay for the Ranked Set Sampling (RSS) subsample of the MMSE data (N = 320), directly corresponding to the component estimates reported in Table 3. The overall density (black solid line) exhibits a multimodal pattern, with the dominant peak driven by Component 3 (red curve), which captures the moderate impairment range with its high proportion of 58.28% and mean of 19.44 (SD = 3.8). A broad shoulder in the severe-to-moderate range is represented by Component 1 (green curve; 23.91%, mean = 10.89, SD = 5.7), contributing substantially to the lower end of the distribution. The mild impairment range is captured by Component 2 (orange curve; 10.49%, mean = 26.15, SD = 1.61), showing a concentrated peak, while Component 4 (purple curve; 7.32%, mean = 29.64, SD = 0.5) represents the near-normal range with tight concentration at the high end. Vertical dashed lines indicate the component means, visually confirming that RSS sampling with age as the auxiliary variable concentrated effort on the central portion of the cognitive spectrum (Components 1 and 3, together comprising over 82% of the sample). However, compared to coreset sampling, RSS substantially undersamples both extremes, particularly the near-normal group (Component 4) and the severely impaired range, highlighting the inherent bias of ranked set sampling toward the middle of the distribution when the auxiliary variable is only moderately correlated with the outcome of interest.

3.3.2. Coreset Sampling (CS)

Table 4 summarizes the four-component Gaussian Mixture Model (GMM) fitted to the coreset subsample of the MMSE data (N = 198, representing a compact, error-controlled subset of the original 663 observations). The coreset approach preserves the multimodal structure effectively, with Component 2 remaining the most prominent (60.35%) and centered in the moderate impairment range (mean = 18.93, SD = 3.7), while Component 3 (19.7%, mean = 25.47, SD = 2.0) and Component 4 (8.38%, mean = 29.43, SD = 0.5) capture the mild-to-near-normal ranges with increasing tightness. Notably, Component 1 (severe impairment) is recovered at 11.57% (mean = 7.64, SD = 4.8), showing substantially

better representation of the lower tail compared to ranked set sampling. Overall, these results demonstrate that coreset sampling successfully approximates the full dataset's distribution using only about 30% of the observations, maintaining good coverage across severity levels with minimal bias.

Table 4
Gaussian Mixture Model Component Estimates Based on Coreset Sampling (N = 198).

Component	Proportion (%)	Mean	SD	Min Range	Max Range
Component 1	11.57	7.64	4.8	0	17.2
Component 2	60.35	18.93	3.7	11.5	26.36
Component 3	19.7	25.47	2	21.58	29.37
Component 4	8.38	29.43	0.5	28.39	30

Figure 6 displays the fitted four-component Gaussian Mixture Model (GMM) density overlay for the coreset subsample of the MMSE data (N = 198), directly corresponding to the component estimates reported in Table 4. The overall density (black solid line) remains clearly multimodal, with a prominent central mode driven by Component 2 (orange curve), which dominates the distribution around the moderate impairment range (15–22), aligning with its high proportion of 60.35% and mean of 18.93 (SD = 3.7) in Table 4. A distinct secondary peak in the mild-to-near-normal range (25–30) is captured by Component 3 (red curve; 19.7%, mean = 25.47) and especially Component 4 (purple curve; 8.38%, mean = 29.43, very low SD = 0.5), showing tight concentration at the high end. Component 1 (green curve; 11.57%, mean = 7.64) contributes a visible but smaller tail at the severe low end (0–10), demonstrating substantially better recovery of the lower tail compared to ranked set sampling. Vertical dashed lines approximate the component means, visually confirming that coreset sampling preserves the full multimodal structure of the original MMSE distribution remarkably well despite using only about 30% of the data, with balanced representation across severe, moderate, mild, and near-normal severity levels.

3.4. Overlap Percentages of Gaussian Mixture Model Components: Full vs. Subsampled Datasets

Table 5 compares the pairwise overlap percentages between the four Gaussian Mixture Model (GMM) components across the full original dataset (N = 663), Ranked Set Sampling (RSS, N = 320), and Coreset sampling (N = 198).

The full dataset shows the highest overlaps, particularly between adjacent components (e.g., 46.3% for Components 1 & 2, 27.6% for 2 & 3), reflecting the naturally continuous and overlapping nature of MMSE scores in the Alzheimer’s cohort. Both subsampling methods reduce overlaps significantly compared to the full GMM. Still, in different ways, RSS achieves the lowest overlaps overall (e.g., 3% for 1 & 2), though it increases overlap in higher components (35.7% for 1 & 3, 19.3% for 2 & 3) due to heavy concentration in the moderate-to-mild range. Coreset sampling provides a balanced reduction (e.g., 18.23% for 1 & 2,

23.17% for 2 & 3, near-zero for distant pairs like 1 & 4), preserving realistic transitional continuity while maintaining clearer separation at the extremes and better recovering the severe impairment component.

Table 5
Overlap Percentages Between Gaussian Mixture Model Components Using Rank Set and Coreset Sampling.

Components	Gaussian Mixture Model	Rank set sampling	Coreset sampling
	Overlap (%)	Overlap (%)	Overlap (%)
Component 1 VS Component 2	46.3	3	18.23
Component 1 VS Component 3	9	35.7	0.73
Component 1 VS Component 4	5.7	0.1	0
Component 2 VS Component 3	27.6	19.3	23.17
Component 2 VS Component 4	20	8.1	0.89
Component 3 VS Component 4	1.6	1.2	8.69

Overall, Coreset sampling offers the best trade-off: it minimizes excessive overlap, closely approximates the true severity distribution and component means, and delivers the most faithful GMM representation using only about 30% of the original data.

4. Discussion

The main aim of this paper was to use a four-component Gaussian Mixture Model (GMM) and describe the multimodal and continuous distribution of Mini-Mental State Examination (MMSE) scores in an Alzheimer's disease (AD) sample (N = 663) and carefully consider Ranked Set Sampling (RSS; N = 320) and Coreset sampling (N = 198) as effective subsampling methods. These methods enable a computationally practical and accurate approximation of the full dataset multimodal structure- capturing important characteristics of the multimodal in terms of component proportions, means, and variances, with very low sample size and processing cost. One of the major successes is that these subsampling methods also result in reduced overlap among the components, which in turn produces better separation of the cognitive subgroups (e.g., severe-to-moderate, moderate, mild, and near-normal impairment), which in turn increases the interpretability and reliability of the identified AD severity profiles.

The entire GMM showed significant pairwise overlap among neighboring items (e.g., 46.5% between Components 1 and 2 of severe-to-moderate impairment), which is compatible with the non-

homogeneous and graded cognitive deterioration in AD (40). RSS overweighted the distribution around moderate impairment (Component 3: 58.28, mean = 16.95), which gives it overlaps (e.g., 35.7% between 1 and 3, almost zero between pairs between severe Component 1) but does not recover the extremes very well (32). On the contrary, Coreset sampling maintained a balanced structure (severe: 11.61, mean = 7.67; moderate: 60.29, mean = 18.93) with moderate overlaps (e.g., 18.23% between 1 and 2) and gave an overall weighted mean (~ 19.79) that is nearest to the true value (19.74) and more consistent with clinically defined severity (41).

These results are consistent with previous studies that use GMMs to analyze AD-related cognitive data, in which mixture models have outlined severity subtypes based on neuroimaging and neuropsychological data, and in many cases, the distributions overlap, indicative of continuous progression (42, 43). The idea of subsampling comparisons can be related to the literature on the effectiveness of GMM in large-scale data: coresets have been demonstrated to give provable approximations of full GMM fits with size-independent error rates, and to give better component structure and means in multimodal data (44–46). RSS outcomes are reflective of recorded biases in unbalanced medical data, with ranked methods focusing on common ranges (e.g., moderate impairment) and underrepresenting tails, like the difficulties in ADNI proteomics and MRI analyses (32). Our study has an overlap between metrics calculated based on the integral of pointwise minimum densities, which are extensions of quantitative continuity measures, used in the clustering of cognitive impairments, to support transitional interpretations of AD progression (42, 44).

In clinical terms, lower but informative overlaps in the coreset GMM increase subtype separability and improvement of progression risk strata, and there is an implication concerning enriching trials through baseline MMSE to predict the decline rates (47). The high fidelity of Coreset, which can be used to establish a near-true means with only 30% data, makes it an effective instrument of scalable biomarker modeling in AD, in particular in resource-constrained environments (48).

Due to the assumptions of Gaussian components (which are unlikely to describe the non-normal tails of MMSE distributions) and the relatively small size of the full dataset (N = 663), this study has its limitations, as it is difficult to generalize the findings to larger and more heterogeneous samples. It is important to note that coreset algorithm is especially beneficial in large data settings, where it scales approximations of high fidelity with low computational expense, and the error bound does not depend on the sample size; even in our small AD dataset, Coreset was more successful than RSS in mean recovery and balanced representation, but it would be more successful in large-scale AD research. The framework could be expanded in future work to include multimodal biomarkers or hybrid machine learning to achieve greater robustness in personalized AD prediction.

5. Conclusion

This paper used a four-component Gaussian Mixture Model to describe Mini-Mental State Examination (MMSE) results in an Alzheimer's disease cohort to show that there is a multimodal but continuous

distribution of cognitive impairment with significant overlap between adjoining severity levels. This observation promotes the idea of thinking about the progression of Alzheimer's disease as a continuum and not categorical stages, and it has significant clinical staging implications and implications for the design of trials. We considered two sampling strategies based on their effectiveness and representativeness. Using age as an auxiliary variable, Ranked Set Sampling focused its effort to moderate the range of impairment, but did not fully represent both ends. Coreset sampling, with a fraction of the original data, on the other hand, retained the multimodal structure of the distribution, equal representation of all severity levels, and was able to recover the severe impairment component with clinically meaningful overlaps. Coreset sampling was found to be better with respect to the approximation of the entire GMM structure, providing a computationally efficient route to analyze large datasets. These results help to justify continuous measures in predictive modeling and confront the hard categorical cut-offs in clinical trials. These results need to be confirmed by future studies in larger cohorts, and multimodal biomarkers, such as neuroimaging and genetic variables, should be incorporated in future research to complement individualized intervention in the study of Alzheimer's disease.

Declarations

Ethical Approval and Consent to Participate: This study was conducted in accordance with the principles of the Declaration of Helsinki (2013). Ethical approval was obtained from the Institutional Review Board of the National Institute of Neurosciences and Hospital, Dhaka, Bangladesh (Approval No: NINS/2023/312). Written informed consent was obtained from all participants. For participants with severe cognitive impairment, informed consent was provided by legally authorized representatives, with assent from participants wherever feasible.

Consent for Publication: Not applicable. This manuscript does not contain any person's data in any form (including individual details, images, or videos).

Availability of Data and Materials: The datasets generated and analyzed during the current study are not publicly available due to patient privacy and institutional data protection policies, but are available from the corresponding author on reasonable request and with appropriate ethical approval. All R code used for the Gaussian Mixture Models, subsampling, and model evaluation is publicly available on GitHub: https://github.com/Sakther12/MMSE_Score_R_Code/blob/main/R_code_dementia.R

Competing Interests: The authors declare that they have no competing interests, financial or otherwise, that could have influenced the work reported in this paper.

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions:

Sharmin Akther: Data curation, Formal analysis, Investigation, Software, Validation, Visualization, Methodology, Writing – original draft, Writing – review & editing (major), Formatting, and final editing.

Azizur Rahman: Conceptualization, Project administration, Formatting and Final editing, Supervision.

Rumana Rois: Review & editing.

All authors have read and agreed to the published version of the manuscript.

Acknowledgments: The authors gratefully acknowledge the National Institute of Neurosciences and Hospital, Dhaka, Bangladesh, for providing access to the clinical data. We thank all participants and their families for their cooperation in this study.

References

1. Nichols, E. et al. Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: an analysis for the Global Burden of Disease Study 2019. *Lancet Public. Health*. **7** (2), e105–e125. 10.1016/S2468-2667(21)00249-8 (2022).
2. Livingston, G. et al. Dementia prevention, intervention, and care: 2020 report of the < em>Lancet Commission. *Lancet* **396** (10248), 413–446. 10.1016/S0140-6736(20)30367-6 (2020).
3. Naheed, A. et al. Prevalence of dementia among older age people and variation across different sociodemographic characteristics: a cross-sectional study in Bangladesh. *Lancet Reg. Health - Southeast. Asia*. **17**, 100257. 10.1016/j.lansea.2023.100257 (2023).
4. Kabir, Z. N. & Herlitz, A. The Bangla Adaptation of Mini-mental State Examination (BAMSE): an instrument to assess cognitive function in illiterate and literate individuals. *Int. J. Geriatr. Psychiatry*. **15** (5), 441–450. 10.1002/(SICI)1099-1166(200005)15:5%3C441::AID-GPS142%3E3.0.CO;2-O (2000).
5. Ahmed, S. M., Hossain, M. A., RajaChowdhury, A. M. & Bhuiya, A. U. The health workforce crisis in Bangladesh: shortage, inappropriate skill-mix and inequitable distribution. *Hum. Resour. Health*. **9** (1), 3. 10.1186/1478-4491-9-3 (2011).
6. Folstein, M. F., Folstein, S. E. & McHugh, P. R. Mini-mental state. *J. Psychiatr Res.* **12** (3), 189–198. 10.1016/0022-3956(75)90026-6 (1975).
7. Mitchell, A. J. A meta-analysis of the accuracy of the mini-mental state examination in the detection of dementia and mild cognitive impairment. *J. Psychiatr Res.* **43** (4), 411–431. 10.1016/j.jpsychires.2008.04.014 (2009).
8. Arevalo-Rodriguez, I. et al. Mini-Mental State Examination (MMSE) for the early detection of dementia in people with mild cognitive impairment (MCI). *Cochrane Database Syst. Reviews*. **2021** (7). 10.1002/14651858.CD010783.pub3 (2021).
9. Creavin, S. T. et al. Mini-Mental State Examination (MMSE) for the detection of dementia in clinically unevaluated people aged 65 and over in community and primary care populations. *Cochrane*

Database Syst. Reviews. **2016** (4). 10.1002/14651858.CD011145.pub2 (2016).

10. Ferreira, D., Nordberg, A. & Westman, E. Biological subtypes of Alzheimer disease. *Neurology* **94** (10), 436–448. 10.1212/WNL.0000000000009058 (2020).
11. Graff-Radford, J. et al. New insights into atypical Alzheimer's disease in the era of biomarkers. *Lancet Neurol.* **20** (3), 222–234. 10.1016/S1474-4422(20)30440-3 (2021).
12. Fraley, C. & Raftery, A. E. Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc.* **97** (458), 611–631 (2002).
13. Nettiksimmons, J., DeCarli, C., Landau, S. & Beckett, L. Biological heterogeneity in ADNI amnestic mild cognitive impairment. *Alzheimer's Dement.* **10** (5), 511. 10.1016/j.jalz.2013.09.003 (2014).
14. Leoutsakos, J. S. et al. Latent classes of course in Alzheimer's disease and predictors: the Cache County Dementia Progression Study. *Int. J. Geriatr. Psychiatry*. **30** (8), 824–832 (2015).
15. Haixiang, G. et al. Learning from class-imbalanced data: Review of methods and applications. *Expert Syst. Appl.* **73**, 220–239. 10.1016/j.eswa.2016.12.035 (2017).
16. Johnson, J. M. & Khoshgoftaar, T. M. Survey on deep learning with class imbalance. *J. Big Data*. **6** (1), 27. 10.1186/s40537-019-0192-5 (2019).
17. McLachlan, G. & Peel, D. *Finite Mixture Models* (Wiley, 2000). 10.1002/0471721182
18. Bachem, O., Lucic, M. & Krause, A. Practical coreset constructions for machine learning. arXiv preprint arXiv:170306476. (2017).
19. Feldman, D., Faulkner, M. & Krause, A. Scalable training of mixture models via coresets. *Adv. Neural Inf. Process. Syst.* ;**24**. (2011).
20. Braverman, V., Feldman, D., Lang, H., Statman, A. & Zhou, S. New frameworks for offline and streaming coreset constructions. arXiv preprint arXiv:161200889. (2016).
21. McIntyre, G. A method for unbiased selective sampling, using ranked sets. *Aust J. Agric. Res.* **3** (4), 385–390. 10.1071/AR9520385 (1952).
22. Chen, Z., Bai, Z. & Sinha, B. K. *Ranked Set Sampling* Vol. 176 (Springer New York, 2004). 10.1007/978-0-387-21664-5
23. McLachlan, G. & Peel, D. *Finite Mixture Models* (Wiley, 2000). 10.1002/0471721182
24. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. mixtools: An *R* Package for Analyzing Finite Mixture Models. *J. Stat. Softw.* **32** (6). 10.18637/jss.v032.i06 (2009).
25. Hatefı, A., Jafari Jozani, M. & Ziou, D. Estimation and classification for finite mixture models under ranked set sampling. *Stat. Sin*. 10.5705/ss.2012.178 (2014).
26. McIntyre, G. A method for unbiased selective sampling, using ranked sets. *Aust J. Agric. Res.* **3** (4), 385–390. 10.1071/AR9520385 (1952).
27. Feldman, D., Faulkner, M. & Krause, A. Scalable training of mixture models via coresets. *Adv. Neural Inf. Process. Syst.* ;**24**. (2011).
28. Lucic, M., Faulkner, M., Krause, A. & Feldman, D. Training gaussian mixture models at scale via coresets. *J. Mach. Learn. Res.* **18** (160), 1–25 (2018).

29. Hakim, M. et al. Correlation of Cognitive Status and Atrophy Score in Alzheimer's Disease Among the Bangladeshi Population. Cureus. Jul **31**. (2024). 10.7759/cureus.65833
30. Tombaugh, T. N. & McIntyre, N. J. The Mini-Mental State Examination: A Comprehensive Review. *J. Am. Geriatr. Soc.* **40** (9), 922–935. 10.1111/j.1532-5415.1992.tb01992.x (1992).
31. Folstein, M. F., Folstein, S. E. & McHugh, P. R. Mini-mental state. *J. Psychiatr Res.* **12** (3), 189–198. 10.1016/0022-3956(75)90026-6 (1975).
32. Chen, Z., Bai, Z. & Sinha, B. K. *Ranked Set Sampling* Vol. 176 (Springer New York, 2004). 10.1007/978-0-387-21664-5
33. McIntyre, G. A method for unbiased selective sampling, using ranked sets. *Aust J. Agric. Res.* **3** (4), 385–390. 10.1071/AR9520385 (1952).
34. Feldman, D., Faulkner, M. & Krause, A. Scalable training of mixture models via coresets. *Adv. Neural Inf. Process. Syst.* ;**24**. (2011).
35. Braverman, V., Feldman, D., Lang, H., Statman, A. & Zhou, S. New frameworks for offline and streaming coreset constructions. arXiv preprint arXiv:161200889. (2016).
36. Bachem, O., Lucic, M. & Krause, A. Practical coreset constructions for machine learning. arXiv preprint arXiv:170306476. (2017).
37. Bachem, O., Lucic, M. & Krause, A. Scalable k -Means Clustering via Lightweight Coresets. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York, NY, USA: ACM; pp. 1119–27. (2018). 10.1145/3219819.3219973
38. McLachlan, G. & Peel, D. *Finite Mixture Models* (Wiley, 2000). 10.1002/0471721182
39. Benaglia, T., Chauveau, D., Hunter, D. R. & Young, D. mixtools: An *R* Package for Analyzing Finite Mixture Models. *J. Stat. Softw.* **32** (6). 10.18637/jss.v032.i06 (2009).
40. Li, R. et al. Gaussian Mixture Models and Model Selection for [18F] Fluorodeoxyglucose Positron Emission Tomography Classification in Alzheimer's Disease. *PLoS One*. **10** (4), e0122731. 10.1371/journal.pone.0122731 (2015).
41. McLachlan, G. & Peel, D. *Finite Mixture Models* (Wiley, 2000). 10.1002/0471721182
42. Forouzannezhad, P. et al. A Gaussian-based model for early detection of mild cognitive impairment using multimodal neuroimaging. *J. Neurosci. Methods*. **333**, 108544. 10.1016/j.jneumeth.2019.108544 (2020).
43. Perneczky, R. et al. Mapping Scores Onto Stages: Mini-Mental State Examination and Clinical Dementia Rating. *Am. J. Geriatric Psychiatry*. **14** (2), 139–144. 10.1097/01.JGP.0000192478.82189.a8 (2006).
44. Kennedy, R. E., Cutter, G. R., Wang, G. & Schneider, L. S. Using baseline cognitive severity for enriching Alzheimer's disease clinical trials: How does Mini-Mental State Examination predict rate of change? *Alzheimer's Dementia: Translational Res. Clin. Interventions*. **1** (1), 46–52. 10.1016/j.trci.2015.03.001 (2015).

45. Dubey, R. et al. Analysis of sampling techniques for imbalanced data: An n = 648 ADNI study. *Neuroimage* ;**87**:220–241. doi:10.1016/j.neuroimage.2013.10.005. (2014). PubMed PMID: 24176869.
46. Lucic, M., Faulkner, M., Krause, A. & Feldman, D. Training Gaussian Mixture Models at Scale via Coresets [Internet]. Jan 15. (2018). Available from: http://arxiv.org/abs/1703.08110
47. Gomes, C. C., Boisvert, J. & Deutsch, C. V. Gaussian Mixture Models Learning Objectives • Review the theory and implementation of Gaussian Mixture Models (GMM) • Understand the application of GMMs in Geostatistics • Demonstrate application of the GMM with practical examples. Report.
48. Nowakowska, E., Koronacki, J. & Lipovetsky, S. Tractable Measure of Component Overlap for Gaussian Mixture Models [Internet]. Jul 27. (2014). Available from: http://arxiv.org/abs/1407.7172

Figures

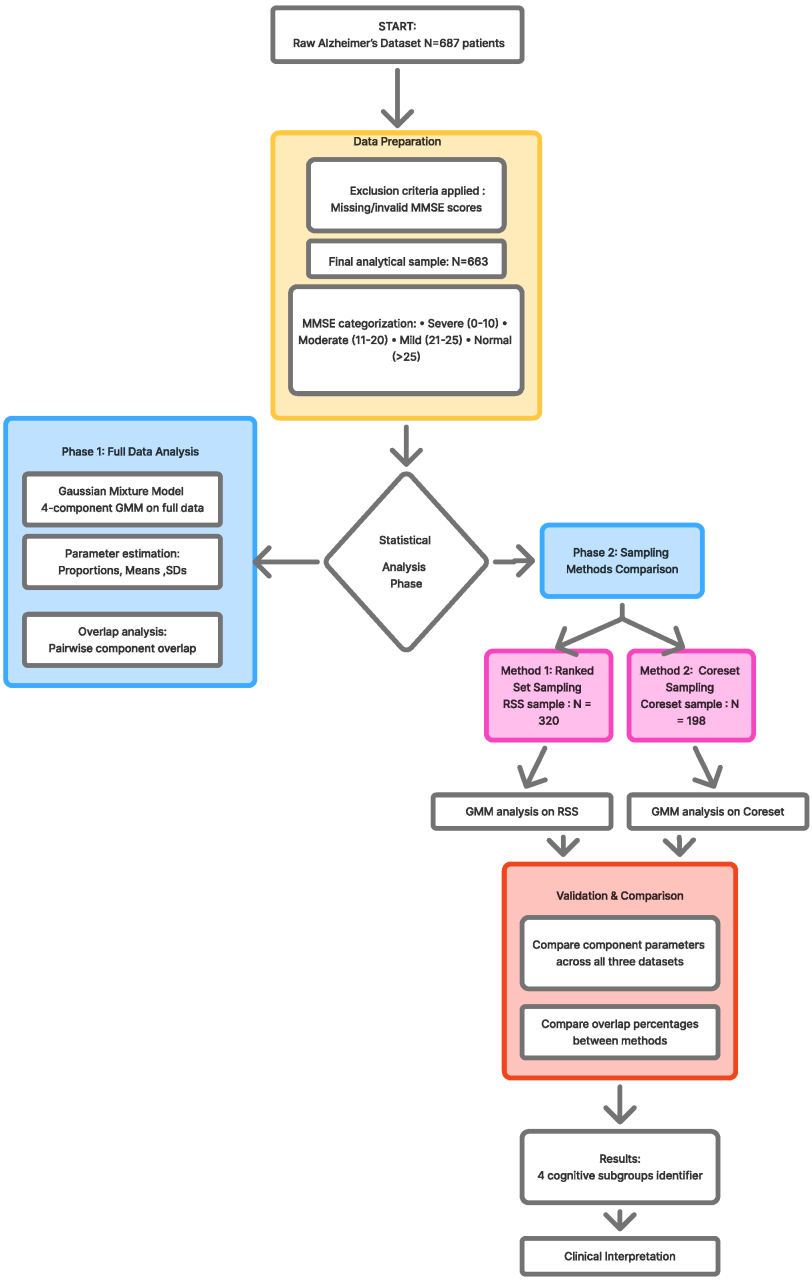


Figure 1

Analytical framework for identifying cognitive subgroups: From raw MMSE data to Gaussian mixture modeling using Ranked Set Sampling and coreset approaches in Alzheimer's disease cohort.

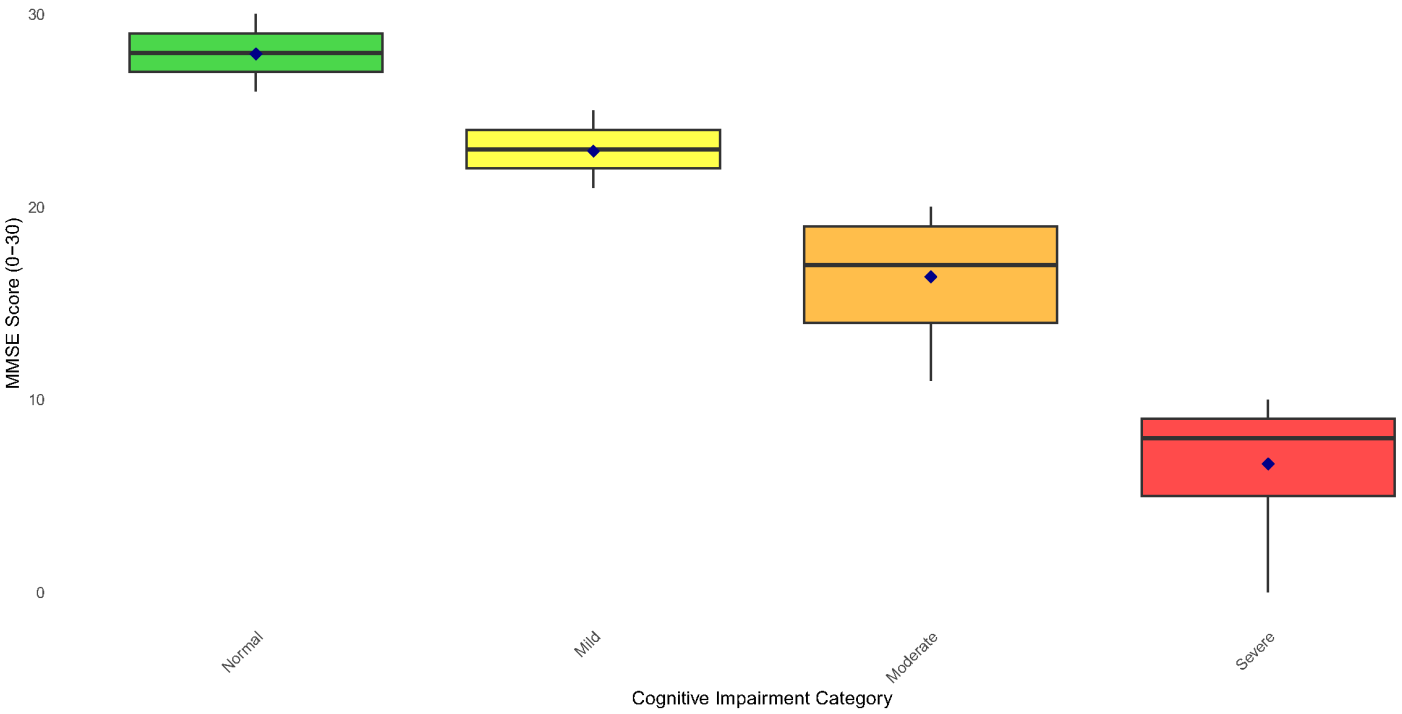


Figure 2

Boxplot of Mini-Mental State Examination Scores Stratified by Cognitive Impairment Severity in Alzheimer's Disease.

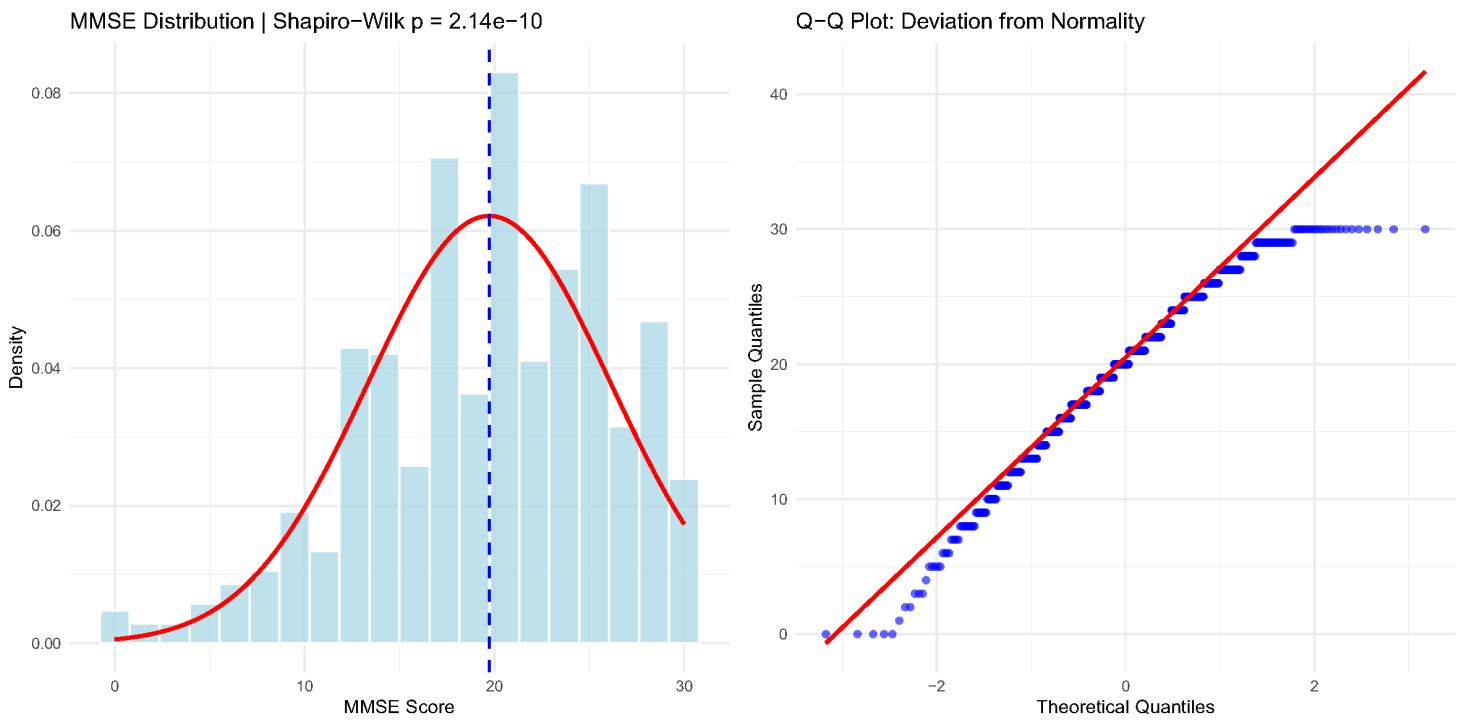


Figure 3

Multimodal Distribution and Departure from Normality of MMSE Scores in Alzheimer's Disease.

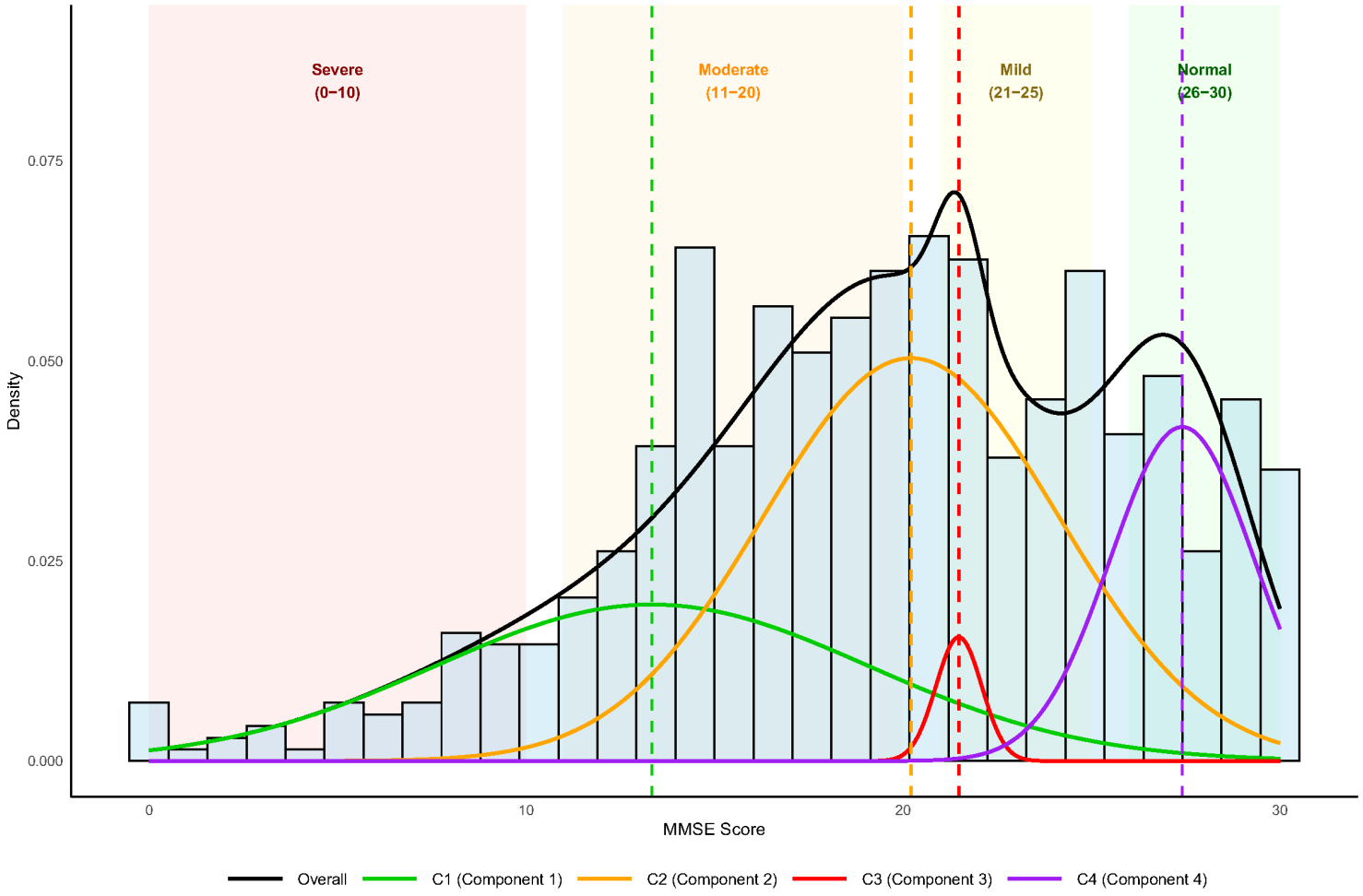


Figure 4

Four-Component Gaussian Mixture Model Fitted to the Full MMSE Dataset (N = 663).

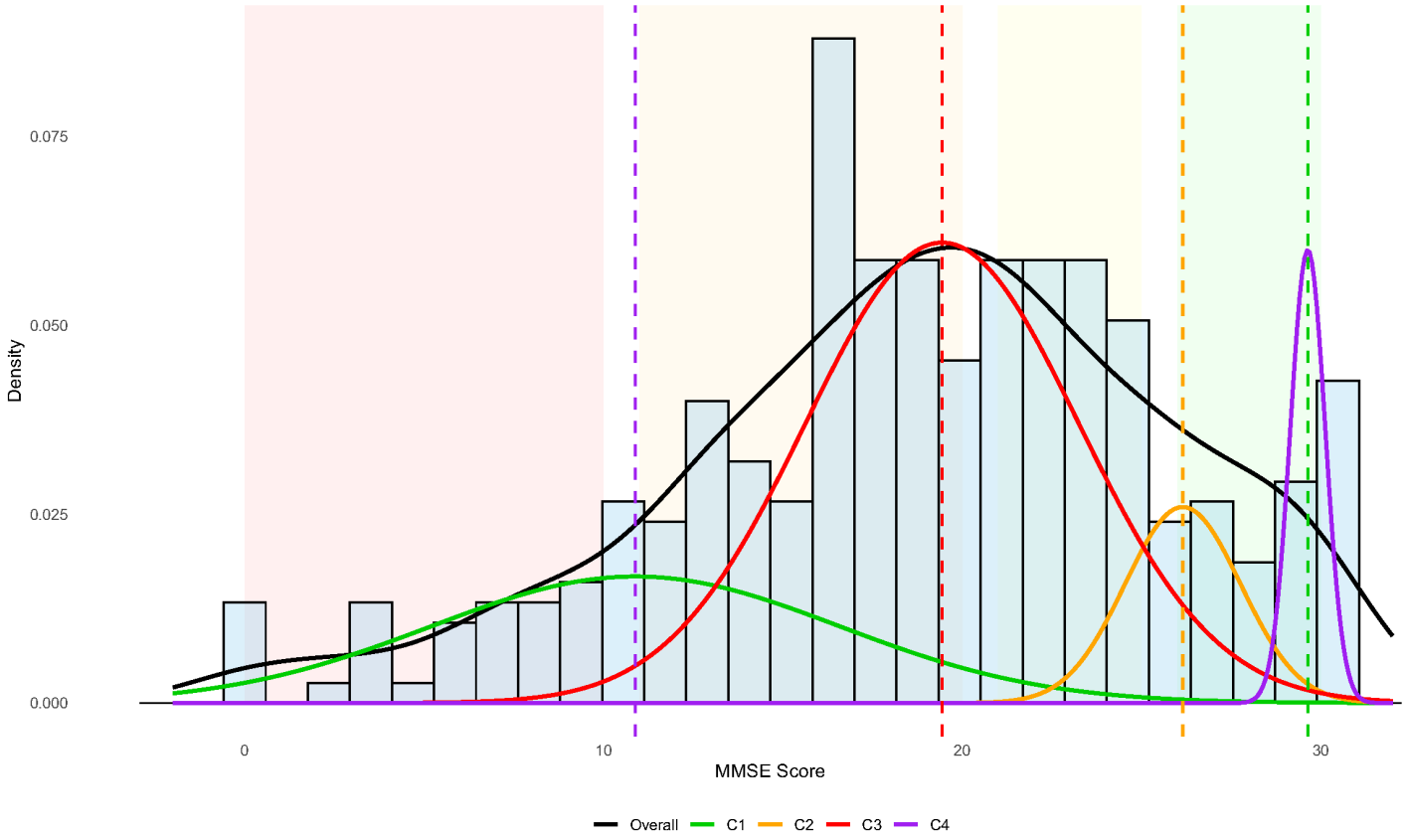


Figure 5

Fitted Four-Component Gaussian Mixture Model to the Ranked Set Sampling MMSE Subset (N = 320).

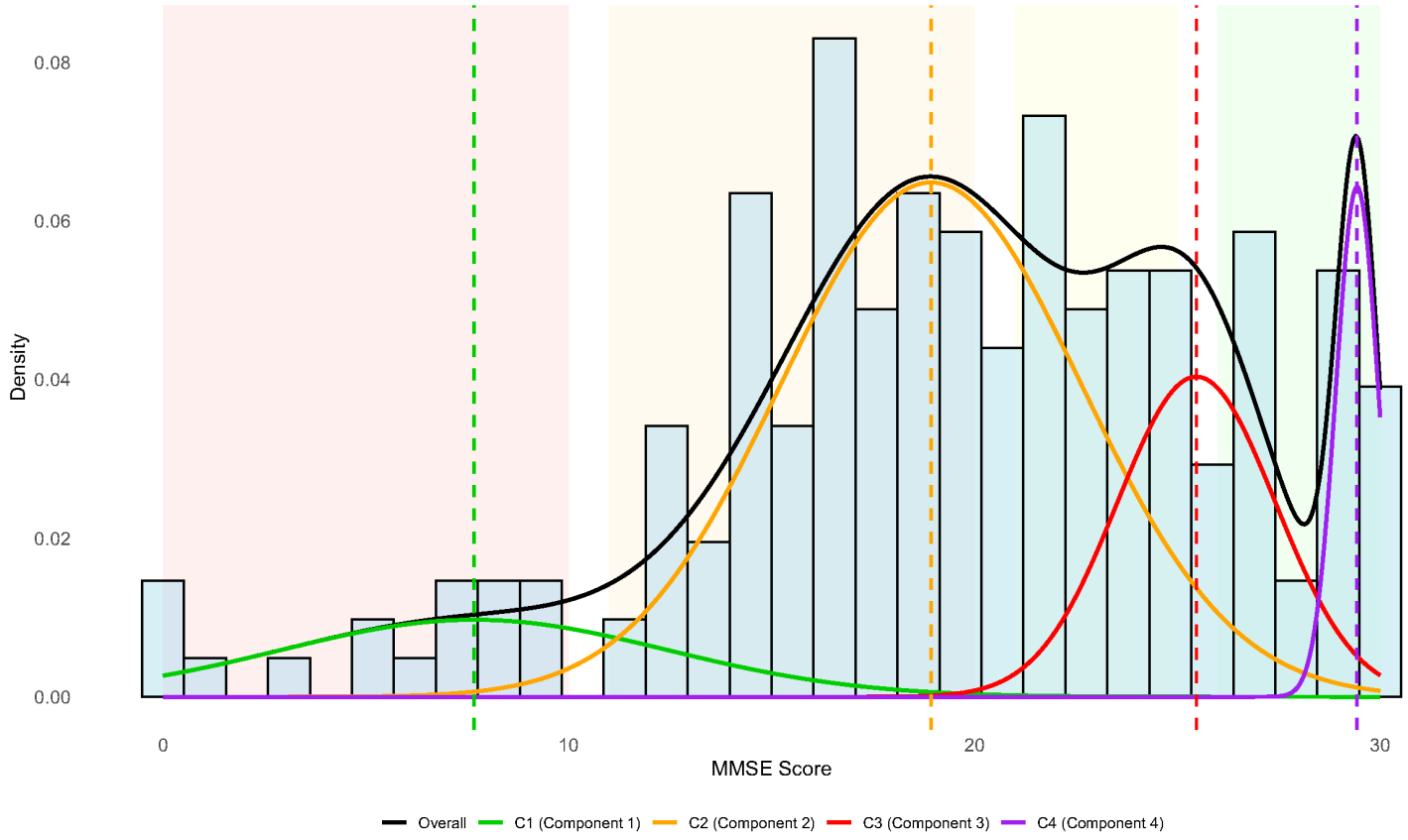


Figure 6

Multimodal Density Fit and Component Contributions in the Coreset Subsample of MMSE Scores (N = 198).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile.docx