# Supplement: Cluster Analysis - Further Methods and Results

Cluster Analysis Methods:

Two unsupervised machine learning clustering algorithms were assessed: k-means and hierarchical clustering, to determine the optimal algorithm that fit the data. Identification of the optimal clustering algorithm between these two as well as the optimal number of cluster (K) for each algorithm was evaluated using the *clValid()* function from the [clValid] R package. This function assessed the internal (i.e. connectivity, silhouette coefficient, Dunn index) and stability measures (average proportion of non-overlap, average distance, average distance between means, figure of merit) of each clustering algorithm. Complementary cluster number optimisation methods were used for the outperforming clustering algorithm according to the *clValid()* function; these include the elbow method and the gap statistic method (both using the *fviz\_nbclust()* function from the [factoextra] R package) as well as using the *NbClust()* function from the R package [NbClust] which comprised 30 indices for determining the optimal numbers of clusters.

For hierarchical clustering approach, pairwise distance between patients was measured using the Euclidean distance, with the following linkage methods being assessed: complete, average, single, Ward D, and Ward D2. Validation of the hierarchical clustering has been assessed using the correlation between the Euclidean distances and the cophenetic distances (i.e. a measure of pairwise distance within the hierarchical cluster tree) using the *cor()* function.

Cluster Analysis Results:

|  |
| --- |
| Supplementary Figure 1: Pairwise dissimilarity matrix |
| A screenshot of a computer  Description automatically generated |
| The cross-sectional data set, comprising 94 patients with SLE for whom 10 different cognitive tests were assessed, was considered clusterable based on Hopkins statistics value of 0.64 and the ordered pairwise dissimilarity matrix |

|  |
| --- |
| Supplementary Figure 2: Cluster optimisation methods |
| A) Elbow Method  A screenshot of a computer  Description automatically generated with medium confidence |
| B) Silhouette Method  A screenshot of a computer  Description automatically generated |
| C) NbClust function  Graphical user interface, chart  Description automatically generated |
| D) Gap Statistic Method  A screenshot of a computer  Description automatically generated with medium confidence |
| An optimal number of two clusters (K = 2) for k-means clustering was further confirmed by the elbow method (A), the silhouette method (B) and the NbClust function (C), albeit not by the gap statistics (D) (average silhouette width: 0.26; Dunn index: 0.22) |