Use of clustering techniques for clinical and epidemiological research: practical tips using an example from rheumatology

Jansen, M.; Bakker, M.M.; Sepriano, A.R.; Shkedy, Z.; Landewe, R.L.; Ramiro, S.; ... ; Putrik, P.

Journal of Clinical Epidemiology 192 (2026) 112183

COMMENTARY

Use of clustering techniques for clinical and epidemiological research: practical tips using an example from rheumatology

Martine Jansen[a,b], Mark M. Bakker[c,d,e,*], Alexandre R. Sepriano[f,g], Ziv Shkedy[a], Robert L. Landewé[h,i], Sofia Ramiro[f,i], Annelies Boonen[c,d], Polina Putrik[j,k]

[a]*Center for Statistics (CenStat), Data Science Institute (DSI), Hasselt University, Diepenbeek, Belgium*
[b]*Marketing, Communication & PR, Fontys University of Applied Sciences, Eindhoven, The Netherlands*
[c]*Department of Rheumatology, Maastricht University Medical Centre +, Maastricht, The Netherlands*
[d]*Department of Rheumatology, CAPHRI Care and Public Health Research Institute, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands*
[e]*WHO Collaborating Centre for Public Health Leadership & Workforce Development, Department of International Health, CAPHRI Care and Public Health Research Institute, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands*
[f]*Department of Rheumatology, Leiden University Medical Centre, Leiden, The Netherlands*
[g]*Nova Medical School, Universidade Nova de Lisboa, Lisboa, Portugal*
[h]*Rheumatology & Clinical Immunology, Amsterdam University Medical Center, Amsterdam, The Netherlands*
[i]*Department of Rheumatology, Zuyderland Medical Center, Heerlen, The Netherlands*
[j]*Department of Social Medicine, CAPHRI Care and Public Health Research Institute, Faculty of Health, Medicine and Life Sciences, Maastricht University, Maastricht, The Netherlands*
[k]*Living Lab Public Health Mosa, Department of Knowledge & Innovation, Public Health Service South Limburg (GGD Zuid Limburg), Heerlen, The Netherlands*

Accepted 5 February 2026; Published online 10 February 2026

Abstract

Clinical and epidemiological researchers may want to identify or determine groups with similar characteristics that exist within a dataset or population. For this purpose, clustering techniques can be applied. However, clustering can be complex, and results can be misleading if the methods are not applied with fidelity.

The aim of this paper is to provide practical guidance to (clinical) researchers in rheumatology and other fields who wish to explore or confirm (sub) groups within their study population, and are considering to apply clustering techniques to (patient) data. We discuss when clustering is useful for addressing your research question (and when it is not), the need to define the cluster concept, the choice of distance measure between 2 observations, the preprocessing of the data, the assessment of clustering tendency, the types of clustering methods, the determination of the number of clusters and end with the evaluation of the derived clustering solutions using visualizations and statistical measures.

The following methods are presented in this paper: K-means, partitioning around medoids, hierarchical clustering, Density-Based Spatial Clustering of Applications with Noise, spectral clustering, fuzzy clustering, and latent class analysis. To illuminate the different considerations involved in clustering, we use a case example: secondary data from 895 people with rheumatoid arthritis, spondyloarthritis, and gout who completed the Health Literacy Questionnaire (HLQ). In this case example, we compute, appraise and evaluate clustering solutions using different methods in 6 steps, including statistical measures and expert opinion.

The 6 steps proposed in this paper describe a systematic approach to cluster analysis to ensure all important aspects are considered. In the case example, different clustering methods led to different clustering solutions and insights. Qualitative interpretation by content experts was insightful here. Where possible, researchers should apply more than one clustering method fitting to the objective and reflect on eventual differences in solutions.

Keywords: Cluster analysis; Health literacy; Tutorial; R statistical software; Unsupervised learning; Visualization

* Corresponding author. Department of International Health, Faculty of Health, Medicine and Life Sciences, Maastricht University, Duboisdomein 30, 6229 GT, Maastricht, The Netherlands.
E-mail address: m.bakker@maastrichtuniversity.nl (M.M. Bakker).

https://doi.org/10.1016/j.jclinepi.2026.112183

1. Introduction

Clustering refers to a type of method that can search for patterns and insights in data, without knowing upfront what you are looking for. It is a form of unsupervised learning, because there is no known answer (label), with which to check the outcomes of the process. These techniques are generally used for exploratory analysis [1]. This paper focuses on patterns or relationships between observations (representing cases or patients). For this purpose, clustering techniques are useful [1, p 514]. These techniques identify clusters as (previously unknown) subgroups within a set of observations, subgroups of observations that are more similar to each other than to observations in a different subgroup. Other unsupervised learning methods besides clustering exist. For instance, factor analysis or principal component analysis [1, p 496] help look for patterns or relationships between variables (representing measurements, characteristics, or indicators).

Clustering of patients is highly relevant for clinical research. Clustering can help to understand diversity within a patient population, or allow to differentiate treatments between subgroups. For example, a hierarchical cluster analysis was used to identify subgroups by health outcomes and health-care resource utilization among patients with rheumatoid arthritis [2]. Four clusters were detected that differ in disease activity, pain, fatigue, comorbidities and health-care−related costs. Similarly, patients with active systemic lupus erythematosus can be clustered into 3 subgroups, mainly based on T-cell heterogeneity [3].

According to a 2018 systematic review, there has been an increase in studies adopting unsupervised learning in rheumatic and musculoskeletal diseases [4]. Most of the studies in the articles they reviewed used a form of cluster analysis. The review assessed the reporting quality of cluster analyses studies according to a set of criteria, which unfortunately did not include a justification of the clustering method used. Only 2 methods of clustering, hierarchical methods and K-means, were applied in the included studies [4]. Researchers are more likely to apply clustering methods that are frequently reported in publications [5], and thus broader knowledge of available techniques among clinical researchers is important. For example, latent class analysis (LCA) since emerged as a popular clustering method in rheumatology, for example to identify 4 classes (or types) of spondyloarthritis [6,7]. Besides LCA, K-means and hierarchical methods, many more clustering techniques have been proposed in statistical literature, including partitioning around medoids (PAM), density-based spatial clustering of applications (DBSCAN), Spectral clustering, and Fuzzy clustering.

These techniques can also be used to investigate a ground truth or hypothesis on an outcome. The results of clustering will not always confirm those preexisting ideas. This would not mean these were wrong, only that the clustering method found another clustering. When we have found a sensible division in groups (whether from a preexisting idea or via a clustering method), this can be used to classify large cohorts.

The aim of this paper is to give clinical researchers in rheumatology (and other fields) practical and methodological guidance with regard to clustering of patient data via a stepwise, hands-on tutorial, with the help of an example from clinical practice. This paper presents a summary of the key steps and methods in the main text. More technical details, including a simulated dataset, a suggested analysis plan for this type of data and annotated R-code used for calculations are available to the reader as supplementary files. All analyses were performed using R Statistical Software (v4.4.2; R Core Team 2024).

2. Clustering step by step

When considering and performing a cluster analysis, there are several decisions to be made. First, we have to assess if clustering is indeed contributing to our research question (step 1). If so, the next step is to formulate what we consider to be a cluster in our research context, we have to formulate the cluster concept (step 2). Then we have to decide how to evaluate the distance between 2 observations, how can we capture the dissimilarity between 2 observations—for instance, sets of patient data—with a quantitative measure (step 3). After these decisions, we explore the clustering tendency, to get an idea to what extent there are natural clusters in the data. Furthermore, some rules of thumb regarding the required sample size need to be considered, most of which are related to the number of variables used to calculate distances (step 4). Depending on the clustering concept we can apply multiple clustering methods, while deciding on the number of clusters using statistical calculations or expert opinion or both (step 5). Finally (step 6), we have to assess and interpret the clustering solution. Is the clustering solution helpful for the problem at hand?

In the next sections, each of these steps will be briefly discussed. The steps will be illustrated using secondary data from a study on health literacy among people with rheumatoid arthritis, spondyloarthritis, or gout (see Boxes 1−6). Originally, a hierarchical cluster analysis was used to cluster these patients into 10 different profiles of patients differing in strengths and weaknesses across nine health literacy domains [8].

2.1. Step 1: Decide whether clustering is the way to address the research question

The first decision is whether clustering is the best way to answer the research question. If the aim is indeed to find meaningful (sub) groups in the observations (patients), for instance for designing tailored interventions per group, clustering might be the appropriate method. As noted earlier, other techniques for unsupervised learning exist

Box 1 Health Literacy Profiles, the case example for this tutorial

There is increasing recognition of health literacy as a determinant of health that does not represent a unidimensional characteristic. Rather, people vary in their health literacy strengths and weaknesses across domains, resulting in individual challenges. Bakker et al [8] performed a cross-sectional study for which almost 900 patients with rheumatoid arthritis, spondyloarthritis, and gout from 3 hospitals in the Netherlands completed the Health Literacy Questionnaire (HLQ). The HLQ contains 44 items and assesses a person's strengths and difficulties across 9 domains of health literacy ([9], see also Appendix A). If there is some underlying structure in these data that would suggest different health literacy profiles (common patterns of strengths and weaknesses), this could help the clinicians to understand different health literacy needs within a patient population, and allow differentiating treatments between subgroups. This question can be investigated by means of clustering.

besides clustering. The main distinction is between clustering (finding coherent subsets of observations in the data), dimensionality reduction (summarizing a large number of correlated variables into a smaller number of new variables that explain most of the variability in the original set) and association rules (finding if-else patterns in the data) [10]. Note that clustering is an exploratory technique, and often, the observations can be partitioned into meaningful subgroups in more than one way [11]. Moreover, clustering algorithms will almost always produce clusters, which do not necessarily correspond to any clinical reality. Thus, combining statistical techniques with clinical expertise is important (see also Step 6).

2.2. Step 2: Determine cluster concept

What is considered a good type of cluster in the context of the study objective and needs, should be clearly defined in advance. If a cluster is considered as a set of observations that can be represented by a prototype observation, the cluster can be thought of as being a so-called compact shape, with all the observations close together, and separate from other clusters. This is shown in Figure 1A, where the 2 blobs can be clearly distinguished by values on 2 axes, and the prototype per cluster is the center of the blob. Perhaps in some contexts it is reasonable that a cluster is not so compact, as is shown in Figure 1B. Two distinct blobs are visible, where the bottom one is not very compact, but more oblong. A small empty space is visible at the center of the bottom blob. However, interpreting this gap as an indication that the lower point cloud consists of 2 smaller point clouds, is overthinking. This example is generated from a 2-dimensional uniform distribution and the empty space is merely coincidental. It could be plausible that clusters overlap, as shown in Figure 1C, where for some points it is not clear to which point cloud they belong. In a context of text recognition, a cluster could be a character, seen as a series of dots, of which each one is close to one other, but not necessarily all close together, as shown in Figure 1D. The cluster concept is informative of the type of clustering methods to choose in step 6 [8].

2.3. Step 3: Determine distance between observations

The next step is to decide how to capture the dissimilarity between 2 observations, that is how to measure the distance between 2 observations. Instead of a distance measure, sometimes a similarity measure is used. This decision influences the choice of clustering method, as well as the outcome [10].

2.3.1. Choice of variables

When clustering with existing data, it could be that some variables are more interesting to cluster with than other variables. A split can be made between variables that are used for clustering, and other variables, which later could be used in combination with the characteristics of the found clusters to describe or analyze the found clusters.

2.3.2. Types of variables and distance

To calculate a distance between 2 observations, their values need to be compared on the available variables. Variables can be quantitative or qualitative, and are expressed on

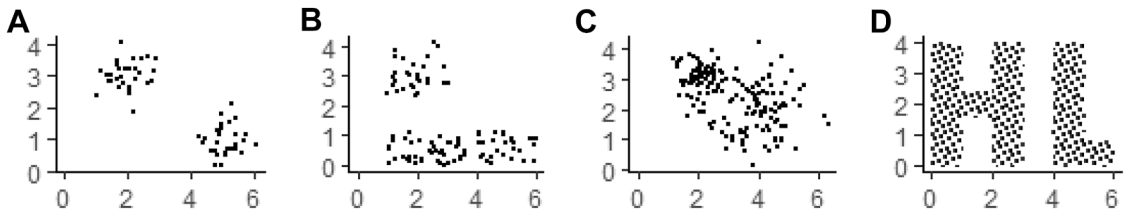

Figure 1. Four different types of cluster concepts. In figure A and B the point clouds are distinct. In figure A they are both compact, while in figure B the bottom one is oblong. Figure C shows a point cloud at the top left that is denser than the point cloud at the right bottom. There even could be some overlap. Figure D shows 2 distinct point clouds that are not convex.

Box 2 Cluster concept for the case example

In the context of health literacy research, a cluster would be a group of patients with a similar health literacy profile (scores on the 9 subscales of HLQ), comparable to plot A above. The differences between patients in the same cluster should be relatively small, and the differences between patients in different clusters should be relatively large, compared to the differences between patients in the same cluster.

different scales (continuous, ordinal scale, categorical). The choice of distance measure depends on the scale of the variables. Distances between quantitative variables are often expressed as Euclidean distance (or squared Euclidean distance).

Variables on an ordinal scale are often transformed to a continuous range between 0 and 1, and from then on treated as continuous variables. For categorical (qualitative) variables the distance can be denoted as 1 in case the values are different for the observations and 0 if they are the same.

As an observation is typically defined by more than just 1 variable, the distance between 2 observations is often seen as the (weighted) sum of the distances between the individual values of the 2 observations over all variables [10]. Other distance measures than Euclidean distance are available. A description of some can be found in Appendix C. A concise table of distances for continuous data including advantages, disadvantages, and indication of processing time can be found in Shirkhorshidi et a [12].

2.3.3. Pre-processing the data or not

Variables usually differ in scale ranges and answer ranges, and therefore differ in variation. In some cases this may be problematic, but it does not have to be. Variation within a variable defines its relative impact on the clustering. A variable with a high variance gets more dominance in determining the distance (and hence the clustering) than a variable with a low variance. A variable with low variance is thus less important for determining the distance, and hence less important for the clustering. Preprocessing influences the weights of the variables taken into the distance measure and because of that also has an impact on some clustering methods [11, p 716].

Consider the example where a distance needs to be established between customers buying clothes based on the variable amount spent in euros and the variable number of items bought. If it is the opinion that both variables are equally important to find customer segments, then a difference of 1 euro is probably not worth as much as a difference of 1 item bought. In these situations, it is better to preprocess both variables, so both variables are on comparable scales. However, if it is the opinion that a difference of 1 in both variables is of a similar importance to determine the distance between customers, preprocessing is not needed.

Sometimes variables have comparable answer ranges, but the given answers show different variation. Depending on the domain specific perception of a cluster, a decision can be made whether the relative differences within a variable are what is important, or the absolute differences. For example: in a survey, one question is about an opinion measured on a continuous scale from 1 to 10. Suppose all respondents would score between 3 and 4. Do we conclude there is not much difference in opinion? That implies we decide it should not have a large effect on the clustering and we look at absolute differences, and hence do not preprocess. If we feel it is important to look at the relative differences, even if all answers are on the low end of the scale, we standardize the data.

There are different standardization methods. Regular standardization scales a variable by subtracting its mean and dividing the remainder by its standard deviation. A method less sensitive to outliers is robust standardization where the median value gets subtracted, and the remainder gets divided by the median absolute deviation. Ordinal variables can be rescaled to the interval [0,1] with borders, or the interval $<0,1>$ without borders [10].

2.4. Step 4: Determine clustering tendency and sample size

Some clustering methods always provide at least 2 clusters, even when there are no actual clusters in the data. This can also be helpful, but it is important to be able to assess whether the obtained clusters are natural clusters, or artifacts of the clustering method, especially when a clustering method results in exactly 2 clusters.

One way of getting an impression about the *clustering tendency* is via the Hopkins statistic [13]. The Hopkins statistic takes a sample of points from the dataset, and calculates for these points the average distance $\overline{dist}_{data}$ to the nearest neighbor. It does the same for an equal sized set

Box 3 Distance for the case example

The HLQ comprises 44 questions scored on a 4- or 5-point ordinal scale addressing 9 different domains of health literacy. Thus, the 9 domains have already been selected as the 9 variables to include in the cluster analysis. We do not yet have a clear opinion on the importance of absolute or relative differences.

To investigate the influence of differences in scales combined with the various types of preprocessing on the resulting clusters, we clustered with the 9 domain variables (obtained via averaging of the 44 questions), with the standardized version, with the robust standardized version and with the translation of the original 44 ordinal questions to a (0,1)-scale and to a [0,1]-scale. Since after preprocessing all variables are on the same continuous scale, we decided to use the (squared) Euclidean distance.

of randomly simulated points. The statistic is calculated by dividing the average distance $\overline{dist}_{random}$ from the random set by the sum of both average distances. If the data set is more or less random, the value of the Hopkins statistic will be around 0.5. If all points in the data set have a very nearby neighbor, the Hopkins statistic will be close to 1, indicating a high clustering tendency. See Appendix B.

There is no general rule for *sample size* regarding clustering. Almost every clustering method will divide a dataset into clusters, whether they are meaningful or not, even when in reality there is no subdivision in clusters. On the other hand, if the sample size is too small, clustering methods can fail to identify an underlying structure, even if there is a true structure in reality [14].

What may be considered an outlier in a particular sample, could be interpreted as a small cluster if the sample size had been larger. When the number of dimensions (variables) increases, the density of the observations can become sparser, so every observation is more like an outlier, with no nearby neighbors. This is called the ''curse of dimensionality'', coined by Bellman in 1961 [10].

There are 2 options to overcome this curse of dimensionality. Either decrease the number of variables, or increase the sample size. One way to reduce the number of variables is by using principal component analysis [10]. Principal Components are the directions in the variable space in which the observations vary the most. If the first 2 Principal Components together explain a high proportion of the total variability, a 2-dimensional plot can be constructed showing how the individual observations relate to these 2 directions. Other options to reduce the number of variables exist, such as nonlinear data reduction, which is not further explored here. For details, consult Lefèvre et al [15].

Several authors have suggested rules of thumb for sample size, often in relation to the number of variables. For instance, Dolnicar et al [14] mention the method of Formann, which states that the sample size for binary data should at least be 2^p, where p is the number of variables. Qiu and Joe suggest at least $10 * p$ per cluster and Dolnicar at least $70 * p$ [both in 14]. Another way of checking if the data space is populated enough (with variables and observations) to do a cluster analysis, is by looking at the previously mentioned clustering tendency. If a clustering tendency is seen, this can be an indicator that the sample size is large enough.

Box 4 Clustering tendency and sample size for the case sample

Some clustering methods will always divide the data in at least 2 clusters, even if no clusters in the data exist. This is an artifact of the clustering method. For the HLQ data, each of the preprocessing variants leads to similar values of clustering tendency, of around 0.755 (Hopkins statistic). This means that if a method would propose 2 clusters, the clusters are probably natural clusters, not an artifact of the method.

The data set from the HLQ contains data from 895 patients, and there are 9 variables. By each of the criteria listed above, the sample size appears to be sufficient.

2.5. Step 5: Choose clustering methods and number of clusters

There are many different clustering methods, see Table 1 for a description of the methods explored in this paper, and

Table 1. Comparing clustering methods with regard to cluster concept and outlier handling

Method	Cluster concept	Outlier handling
K-means	Distinct, nonoverlapping convex clusters, with a center	Sensitive to outliers, can shift cluster centers
Partitioning around medoids	Distinct, nonoverlapping convex clusters with a center	More robust than K-means
Hierarchical clustering	Clusters nested within clusters, clusters higher in the hierarchy not necessarily compact	Sensitive to outliers
Density-based spatial clustering of applications with noise	Distinct clusters according to density, not necessarily convex	Outliers are considered noise, are not assigned to a cluster
Spectral clustering	Distinct, not necessarily convex clusters, adjacent points get clustered together	Robust, gets assigned to cluster
Fuzzy clustering	Observations can be part of more than one cluster at the same time	Sensitive to outliers, can shift cluster centers
Latent class analysis	Assumed is an underlying latent variable i.e. associated with the observed outcomes. Every observation is seen as stemming from a mix of Gaussian distributions, receives for each of these distributions a probability that it stems from that distribution	Sensitive to outliers, could become own cluster

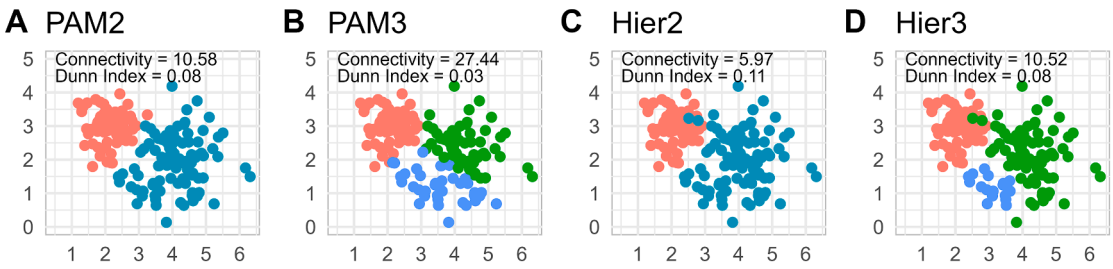

Figure 2. A set of points, clustered with clustering methods PAM and hierarchical clustering, with 2 and with 3 clusters. For each clustering solution the Connectivity and the Dunn Index is calculated. The clustering resulting in the best separated clusters, is the clustering with the lowest connectivity and the highest Dunn index. In this case this is the hierarchical clustering with 2 clusters, show in subplot C. PAM, partitioning around medoids.

Appendix D for more details. Depending on the clustering concept, some clustering methods are a more obvious choice than others. If clusters are seen as being convex and compact (for an example see Figure 1A), the method K-means or the more stable method PAM could be tried. If we believe there is some underlying construct with different classes that could lead for some cases to have similar values on the variables, a method that allows overlap in clusters, such as LCA, is a possibility.

There can be more than one suitable clustering method and each one can help us understand aspects of the data that can be useful. Some of the different clustering solutions can even result in a different number of clusters. In other words, there is not necessarily one unique solution, and the *optimal* number of clusters might be ambiguous.

That said, depending on the clustering algorithm, a statistically optimal number can be obtained in different ways. The elbow method is a frequently method for algorithms that minimize within-cluster variance. The gap statistic is used to formalize this method. For the model-based LCA, a distribution of bootstrapped likelihood ratio test values is used to decide if k + 1 clusters work better than k clusters. Other options for LCA include information criteria or entropy [16].

Besides these more or less objective ways to determine the number of clusters, more subjective (nonstatistical) considerations regarding size, distinctiveness, and practical usefulness can weigh in. This is where steps 5 and 6 are closely linked, as the chosen number of clusters depends on assessment of the clustering solutions described in step 6.

2.6. Step 6: Assess the clustering solutions

The clustering solutions resulting from the previous step must be thoroughly evaluated. The most important aspect of this evaluation is characterizing the clustering: is the clustering helpful, meaningful and useful for answering the research question? Can the individual clusters be clearly described and distinguished? If available, the minimal (clinically) important difference per variable can be considered. Assessors may also consider differences in patterns across variables. Can the researchers interpret and assign meaningful labels (interpretations) to each cluster? Evidently, input of domain experts is crucial here; can the clusters be recognized as distinct groups in practice?

One way to assess the clustering solutions is to visualize them using different figures. Differences in mean value and variance of variables between the clusters can be expressed in tables and plots. One or several domain experts can judge the clustering solutions, possibly supported by a score form. Once the experts have independently assessed the clustering solutions, they can share their views in a group discussion.

Besides the opinion of experts in the domain being studied, there are also statistical measures to describe internal validity and stability or that help comparing clustering solutions. The remainder of this section will address these.

Internal validity measures are useful to describe clusters obtained by partitioning methods that calculate clusters that are compact and well separated. Compact clusters are clusters with a low intracluster variance. Separation can be seen as the intercluster variance. Connectedness describes to what extent observations are placed in the same cluster as their nearest neighbors. Indices that measure internal validity are for instance the Connectivity (the lower the more separated the clusters are), the average Silhouette

Box 5 Clustering methods for the case example

For the HLQ data, using the statistical measures described before, the method DBSCAN resulted in a single cluster with some outliers. Spectral Clustering did not result in useful clusters. Hierarchical clustering on the standardized data resulted in 3 clusters. Other methods resulted in clustering solutions with a statistically optimal number of clusters between 2 (K-means, PAM, Fuzzy) and 9 (LCA). Added to this collection was the originally chosen cluster solution, stemming from a hierarchical clustering method that selected the clinically meaningful clusters. See Appendix D for an overview.

Box 6 Assessing clustering solutions in the case example

For the HLQ data, a number of visualizations were developed to help the content experts assess and label the clustering solutions. The Ophelia Table [20] (with mean scores per domain presented as a heatmap) and the Line Plot give insight into averages per cluster and per variable. The other visuals, the Raster Plot, Bee swarm Plot, Violin Plot and Parallel Coordinates Plot, show the variation per cluster and per variable. See Appendix E - Examples of Visualisations for the HLQ.

Statistical indices indicated that the most stable clusters for our sample HLQ data were obtained with K-means or PAM with 3 clusters on the robust standardized data.

The 4 content experts individually assessed the clustering solutions guided by the visualizations and a structured questionnaire. Some preferred the 2 or 3 cluster solutions, and some were of the opinion that clustering solutions with more clusters would be helpful to inform more tailored care. After these individual assessments, a group discussion took place.

Combining the outcomes of the expert assessment and statistical insights led to the suggestion of a dual approach: Organize the clinic around the 3 groups (stemming from K-means 3-cluster solution on the robust standardized data), and spend part of the resources on the smaller, more specific clusters from one of the cluster analyses resulting in more clusters (Hierarchical clustering with 9 clusters or the more stable K-Means, also with 9 clusters). Of note, hierarchical cluster analysis allows for further exploration of meaningful subclusters as part of the method, rather than just the results at the top of the dendrogram (3 or 9 clusters in our example). This which was not explored in this tutorial paper, but it explains the difference between the cluster solutions presented here and the previously published 10 clusters in Bakker et al [8].

coefficient and the Dunn Index [17]. The Dunn Index is the ratio of the smallest distance between observations in different clusters to the largest distance between clusters. A higher Dunn Index implies more compact and separated clusters. See Figure 2 for a short example.

Stability measures stem from a concept called consensus clustering. Consensus clustering is based on the assumption that if there were underlying true segments in the population, repeated sampling would show the most stability at the true number of segments. Consensus clustering is a method to represent the consensus across multiple runs of one or multiple clustering algorithms with various numbers of clusters. It can be used to determine the optimal number of clusters, and to assess the stability of the clusters. For more details, see Monti et al [18].

Other methods for comparing clustering solutions are described by Wagner [19]. They differentiate between 3 types of measures: counting of pairs of elements, summation of set overlaps and information-theoretical mutual information. The Rand Index is a basic example of counting pairs, it is the percentage of pairs of observations that are either in the same cluster for both clustering solutions, or are in a different cluster for both clustering solutions. It is highly dependent on the number of clusters. The Rand Index led to some derived indices such as the Mirkin Metric, that is indeed a metric in the mathematical sense. The second category, summation of set overlaps tries to match clusters that have a maximum absolute or relative number of same observations (overlap). Examples are the Maximum-Match-Measure and the Van Dongen-Measure. The last category is mutual information. If the uncertainty about the cluster in one clustering solution can be reduced by the knowledge about the subjects cluster in a different clustering solution, there is mutual information.

To prevent overinterpretation of a single clustering solution, multiple clustering methods can be used. The resulting clustering solutions can be assessed, via plots, domain expert opinion, and via appropriate indices to indicate the more stable solutions. Some indices are only suitable for specific types of clusters. For instance, when trying to cluster pixels into letters (as in subplot D in Figure 1), looking at the Dunn Index will not be informative, as different clusters can be very close to each other, while at the same time clusters are not necessarily compact.

In case similar datasets are available, external validation is possible by comparing the resulting clustering solutions.

3. Conclusion

When conducting cluster analysis, researchers must make several critical decisions. Envisioning the cluster concept can help the selection of appropriate methods. Some clustering methods align better with specific cluster concepts than other methods. Choices such as whether to standardize the data and which distance measure to use can influence the resulting clustering solutions. Once a distance measure is selected, checking the clustering tendency can help determine whether the identified are natural clusters or an artifact of the chosen clustering method. Reporting on these decisions when publishing clustering results is recommended. In an ideal scenario, both data and code are published as supplementary material. At minimum, the main text should include the results of the decisions made in the steps and the software, functions, and parameters used for the analysis. A brief reporting checklist is provided in Appendix F. Readers interested in reading more on

specific steps or related methods may want to consult additional literature [10−12,15,18,19,21−23].

By combining statistical indices, visualizations, and importantly, domain expert opinion, researchers can work toward obtaining a well-suited clustering solution. However, caution is essential. The choice of method can significantly impact the results. When possible, we recommend applying multiple clustering methods and critically reflecting on possible differences in the outcomes. Cluster analysis is an exploratory technique and while it can provide valuable insights, its findings should be interpreted with care to avoid overinterpretation.

CRediT authorship contribution statement

Martine Jansen: Writing − review & editing, Writing − original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Conceptualization. **Mark M. Bakker:** Writing − review & editing, Supervision, Resources, Investigation, Conceptualization. **Alexandre R. Sepriano:** Writing − review & editing, Investigation, Conceptualization. **Ziv Shkedy:** Writing − review & editing, Supervision, Conceptualization. **Robert L. Landewé:** Writing − review & editing, Investigation, Conceptualization. **Sofia Ramiro:** Writing − review & editing, Conceptualization. **Annelies Boonen:** Writing − review & editing, Resources, Investigation, Conceptualization. **Polina Putrik:** Writing − review & editing, Investigation, Conceptualization.

Declaration of competing interest

There are no competing interests for any author.

Acknowledgments

This article originated from the master thesis (MSc Biostatistics and Data Science) of the first author. The developers of the HLQ provide a step-by-step manual for hierarchical cluster analysis, enabling anyone, even with limited statistical knowledge, to use the method for a specific purpose. This paper in no way intends to replace this method, but aims to equip researchers with a potentially broader set of tools for clustering.

Supplementary files

Supplementary files related to this article can be found at https://doi.org/10.1016/j.jclinepi.2026.112183.

Data availability

Simulated data and R-code used in this paper are available via: https://doi.org/10.34894/KN6ASP.

References

[1] James G, Witten D, T. H, R. T. Unsupervised learning. An introduction to statistical learning. New York, NY: Springer; 2021. https://doi.org/10.1007/978-1-0716-1418-1_12.
[2] Mars N, Kerola AM, Kauppi MJ, Pirinen M, Elonheimo O, Sokka-Isler T. Cluster analysis identifies unmet healthcare needs among patients with rheumatoid arthritis. Scand J Rheumatol 2021;51(5):355−62. https://doi.org/10.1080/03009742.2021.1944306.
[3] Kubo S, Nakayamada S, Yoshikawa M, Miyazaki Y, Sakata K, Nakano K, et al. Peripheral immunophenotyping identifies three subgroups based on t cell heterogeneity in lupus patients. Arthritis Rheumatol 2017;69(10):1917−2096. https://doi.org/10.1002/art.40180.
[4] Han L, Benseler S, Tyrrell P. Cluster and multiple correspondence analyses in rheumatology: paths to uncovering relationships in a sea of data. Rheum Dis Clin North Am 2018;44(2):349−60. https://doi.org/10.1016/j.rdc.2018.01.013.
[5] Clatworthy J, Buick D, Hankins M, Weinman J, Horne R. The use and reporting of cluster analysis in health psychology: a review. Br J Health Psychol 2005;10(3):311−465. https://doi.org/10.1348/135910705X25697.
[6] Sepriano A, Ramiro S, van der Heijde D, van Gaalen F, Hoonhout P, Molto A, et al. What is axial spondyloarthritis? A latent class and transition analysis in the SPACE and DESIR cohorts. Ann Rheum Dis 2020;79:324−31. https://doi.org/10.1136/annrheumdis-2019-216516.
[7] Bosch P, Sepriano A, Marques M, van der Heijde D, Landewé R, van Lunteren M, et al. Change in different classes of chronic back pain suspicious of axial spondyloarthritis: a latent transition analysis of the SPACE cohort. RMD Open 2024;10:e004584. https://doi.org/10.1136/rmdopen-2024-004584.
[8] Bakker MM, Putrik P, Rademakers J, Laar M, Vonkeman MR, Voorneveld-Nieuwenhuis H, et al. Addressing health literacy needs in rheumatology: which patient health literacy profiles need the attention of health professionals? Arthritis Care Res 2021;73(1):100−9. https://doi.org/10.1002/acr.24480.
[9] Osborne RH, Batterham RW, Elsworth GR, Hawkins M, Buchbinder R. The grounded psychometric development and initial validation of the Health Literacy Questionnaire (HLQ). BMC Public Health 2013;13:658. https://doi.org/10.1186/1471-2458-13-658.
[10] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: Data mining, inference, and prediction. New York, NY: Springer; 2009.
[11] Hennig C. Clustering strategy and method selection. In: Hennig C, Meila M, Murtagh F, Rocci R, editors. Handbook of cluster analysis. New York, NY: Chapman & Hall/CRC; 2015:703−30.
[12] Shirkhorshidi AS, Aghabozorgi S, Wah TY. A Comparison Study on Similarity and Dissimilarity Measures in clustering Continuous Data. PLoS ONE 2015;10(12):e0144059. https://doi.org/10.1371/journal.pone.0144059.
[13] Hopkins B, Skellam JG. A new method for determining the type of distribution of plant individuals. Ann Botany New Series 1954;18(70):213−27. https://doi.org/10.1093/oxfordjournals.aob.a083391.
[14] Dolnicar S, Grün B, Leisch F, Schmidt K. Required sample sizes for data-driven market segmentation analyses in tourism. J Travel Res 2014;53:296−306. https://doi.org/10.1177/0047287513496475.
[15] Lefèvre T, Chariot P, Chauvin P. Multivariate methods for the analysis of complex and big data in forensic sciences. Application to age estimation in living persons. Forensic Sci Int 2016;266:581.e1-e9. https://doi.org/10.1016/j.forsciint.2016.05.014.
[16] Curran P, Bauer D. What's the best way to determine the number of latent classes in a finite mixture analysis?. CenterStat blog; 2021.
[17] Brock G, Pihur V, Datta S, clValid SD. an r package for cluster validation. CRAN; 2021. Available at: https://cran.r-project.org/web/packages/clValid/vignettes/clValid.pdf. Accessed February 27, 2026.
[18] Monti S, Tamayo P, Mesirov J, Golub T. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Machine Learn 2003;52:91−118. https://doi.org/10.1023/A:1023949509487.

[19] Wagner S, Wagner D. Comparing clusterings - an overview. Karlsruhe: Universität Karlsruhe, Fakultät für Informatik; 2007. https://doi.org/10.5445/IR/1000011477.

[20] Beauchamp A, Batterham RW, Dodson S, Astbury B, Elsworth GR, McPhee C, et al. Systematic development and implementation of interventions to OPtimise HEalth LIteracy and Access (Ophelia). BMC Public Health 2017;17:100−9. https://doi.org/10.1186/s12889-017-4147-5.

[21] Lefèvre T, Chauvin P. A general framework for a reliable multivariate analysis and pattern recognition in high-dimensional epidemiological data, based on cluster robustness: a tutorial to enrich the epidemiologists' toolkit. Rev Epidemiol Sante Publique 2015;63(1):9−19. https://doi.org/10.1016/j.respe.2014.12.017.

[22] Den Teuling NGP, Pauws SC, van den Heuvel ER. Clustering of longitudinal data: a tutorial on a variety of approaches. Stat Anal Data Min 2026;19:e70052. https://doi.org/10.1002/sam.70052.

[23] Giordani P, Ferraro MB, Martella F. An introduction to clustering with R. Berlin, Germany: Springer Nature; 2020. https://doi.org/10.1007/978-981-13-0553-5.