Journal of Translational Medicine

RESEARCH

Open Access

A robust clustering strategy for stratification unveils unique patient subgroups in acutely decompensated cirrhosis

Sara Palomino-Echeverria[1], Estefania Huergo[1], Asier Ortega-Legarreta[1], Eva M. Uson Raposo[2], Ferran Aguilar[2], Carlos de la Peña-Ramirez[2], Cristina López-Vicario[2,3], Carlo Alessandria[6], Wim Laleman[7], Alberto Queiroz Farias[8], Richard Moreau[2,9,10,11], Javier Fernandez[2], Vicente Arroyo[2], Paolo Caraceni[12,21], Vincenzo Lagani[13,14,15], Cristina Sánchez-Garrido[2], Joan Clària[2,3,4,5], Jesper Tegner[13,14,22,23], Jonel Trebicka[2,16], Narsis A. Kiani[17,20], Nuria Planell[1,18*†], Pierre-Emmanuel Rautou[9,19*†] and David Gomez-Cabrero[1,13*†]

Abstract

Background Patient heterogeneity poses significant challenges for managing individuals and designing clinical trials, especially in complex diseases. Existing classifications rely on outcome-predicting scores, potentially overlooking crucial elements contributing to heterogeneity without necessarily impacting prognosis.

Methods To address patient heterogeneity, we developed ClustALL, a computational pipeline that simultaneously faces diverse clinical data challenges like mixed types, missing values, and collinearity. ClustALL enables the unsupervised identification of patient stratifications while filtering for stratifications that are robust against minor variations in the population (population-based) and against limited adjustments in the algorithm's parameters (parameter-based).

Results Applied to a European cohort of patients with acutely decompensated cirrhosis ($n=766$), ClustALL identified five robust stratifications, using only data at hospital admission. All stratifications included markers of impaired liver function and number of organ dysfunction or failure, and most included precipitating events. When focusing on one of these stratifications, patients were categorized into three clusters characterized by typical clinical features; notably, the 3-cluster stratification showed a prognostic value. Re-assessment of patient stratification during follow-up delineated patients' outcomes, with further improvement of the prognostic value of the stratification. We validated these findings in an independent prospective multicentre cohort of patients from Latin America ($n=580$).

†Nuria Planell, Pierre-Emmanuel Rautou and David Gomez-Cabrero contributed equally to this work.

*Correspondence:
Nuria Planell
nplanellpic@unav.es
Pierre-Emmanuel Rautou
pierre-emmanuel.rautou@inserm.fr
David Gomez-Cabrero
david.gomez.cabrero@navarra.es

Full list of author information is available at the end of the article

Conclusions By applying ClustALL to patients with acutely decompensated cirrhosis, we identified three patient clusters. Following these clusters over time offers insights that could guide future clinical trial design. ClustALL is a novel and robust stratification method capable of addressing the multiple challenges of patient stratification in most complex diseases.

Keywords Stratification, Clustering, Complex diseases, Cirrhosis, ACLF, Patient heterogeneity, Unsupervised learning

Background

Heterogeneity is a prevalent phenomenon observed in numerous diseases, including various types of cancer [1], autoimmune conditions like multiple sclerosis [2], and diabetes [3]. Substantial interindividual changes in phenotype and pathophysiology within a disease often limit the effectiveness of traditional "one-size-fits-all" medicine approaches. This becomes especially critical in diseases where environmental and lifestyle factors also play a significant role. Acutely decompensated cirrhosis, which refers to the rapid development of overt ascites, overt hepatic encephalopathy, variceal haemorrhage, or any combination of these disorders, which often leads to nonelective admission to the hospital of patients who were previously stable, exemplifies significant inter-individual variability [4, 5]. Acutely decompensated cirrhosis encompasses a range of causes of cirrhosis, including alcohol consumption, metabolic dysfunction, viral hepatitis, genetic disorders, or autoimmune biliary diseases. It is often accompanied by comorbidities, which are neither causes nor consequences of cirrhosis, but they increase mortality [5, 6]. Heterogeneity of cirrhosis can also include various precipitating events such as infection or alcoholic-related hepatitis, diverse clinical presentations like ascites, gastrointestinal bleeding, and hepatic encephalopathy, and multiple possible outcomes such as cancer, liver failure, or death. This clinical heterogeneity poses a considerable challenge as it likely accounts for the diverse responses to treatment and outcomes observed in these patients [7]. Therefore, we reasoned that analysing a large population of patients with acutely decompensated cirrhosis should allow us to develop stratification tools.

A major tool for the characterization of patient heterogeneity is the identification of patient subtypes, also defined as patient stratification. Importantly, the World Health Organization (WHO) has acknowledged patient stratification as a valuable approach for enhancing population health management and providing better-tailored services [8]. In conceptual terms, patient stratification can be described as the process of grouping or clustering patients based on specific characteristics or patterns without relying on labelled data or information about future outcomes [9]. Therefore, contrary to scores developed using classical statistical approaches based on the clinical course, stratification can capture features explaining patients' heterogeneity independently of their association with patient outcomes.

Numerous attempts have been made to identify subgroups within clinical datasets [9–11]. However, the lack of a universally applicable approach poses a significant challenge in the field of clustering analysis. Although there have been advancements beyond the classical k-means and hierarchical clustering methods, no general framework still allows the organization and classification of clustering methodologies in the clinical setting [12]. Instead, many ad-hoc applications have been developed for specific scenarios, but their generalizability is often limited. Horne et al.'s review highlights how certain disease labels, such as asthma, can encompass diverse symptoms and causes [10]. They illustrate this lack of generalization by noting that they found 63 studies utilizing cluster analysis to identify different asthma subtypes based on various clinical data. While there is no global classification, these applications can be grouped based on specific characteristics such as managing missing values, collinearity, or mixed data [11]. For instance, when handling missing data, some methods exclude samples from the analysis, potentially resulting in a loss of statistical power, while others rely on a single imputation, overlooking the potential bias that can be introduced [13]. Highly correlated variables represent a challenge. Some methods exclude them, while others employ dimensionality reduction techniques such as Principal Component (PC) reduction to capture underlying lower-dimensional data patterns [14, 15]. However, both decisions may affect the outcome of the clustering, as sensitivity analyses are rarely conducted. Moreover, indiscriminate feature selection can inadvertently remove informative features along with noisy ones, potentially biasing the results [16]. Furthermore, most clustering methodologies assume the existence of a single stratification, disregarding the possibility of having none or multiple valid alternatives for subgrouping the population [17]. Interestingly, trace-based clustering methodologies have recently emerged to aid in the interpretation and validation of the identified subgroups, often requiring domain knowledge and expert input [18]. Within this technique, the proposal involves tracking elements across clusters generated by different runs of the clustering algorithm to identify stable and informative patterns in the data set.

Additionally, the evaluation of clustering outcomes is an open problem that is based on the quality of the

produced clusters. In the case of unsupervised clustering, where no preliminary classification exists, evaluations are typically referenced against theoretical benchmarks. For instance, when addressing the optimal number of clusters, various quality metrics are available, such as the clustering coefficient [19] or the silhouette index [20], among many others. Importantly, while there is no universal methodology that excels across all scenarios for all data sets, as dictated by the "*no free lunch*" theorem [21], there exist strategies that yield high-quality results [22–24]. Another essential measure—referred to as robustness—lacks a precise definition. Robustness, in general terms, signifies the capacity of a system to withstand changes [25]. In our context, we investigate whether a clustering remains stable when subjected to perturbations. In this work, we considered two types of perturbations: those derived from changes in the population and those arising from changes in the algorithm's parameters. In the case of population-based perturbations, we quantify how a given clustering is influenced by variations in the underlying population. Bootstrapping is one approach to address this scenario [26]. In the case of parameter-based perturbations, we assess the impact of parameter adjustments in the clustering algorithm on the identified clustering [27]. Consider a scenario where a parameter "x" defines our clustering strategy. How different is the resulting clustering when using "x=1" versus "x=1.1"? Here, robustness translates to clusterings that maintain stability even when parameter values shift. For the reader's clarity, we will name the two different robustness criteria: **population-based robustness** and **parameter-based robustness.**

Importantly, there is currently no methodology capable of addressing all the aforementioned scenarios while ensuring both definitions of robustness. To address these challenges comprehensively, we developed ClustALL, a novel framework that robustly identifies patient subgroups by addressing all the previously mentioned challenges and limitations of existing methodologies. We applied ClustALL -as a proof-of-concept- to a large prospective cohort of patients non-electively admitted to the hospital for acutely decompensated cirrhosis.

In this study, ClustALL addressed the stratification challenge within a dataset of patients with acutely decompensated cirrhosis [28], characterized by the presence of missing data, mixed data types, and correlated features. It revealed multiple stratification solutions, with one exhibiting special interest in the clinical context and showing prognostic value. We then validated the reproducibility of this stratification using a separate prospective cohort of patients, affirming ClustALL's robustness and reliability. One further aim of the study was to demonstrate the usability of stratification over the disease course, showcasing its prognostic value. Looking beyond cirrhosis, ClustALL holds promise for broader applications in diverse clinical settings, suggesting its potential to revolutionize patient subgroup identification and improve healthcare management.

Methods

ClustALL framework

Given a set of patients affected by a complex disease with clinical data available, the goal of ClustALL is to identify all the possible alternatives to stratify them that are robust and consistent, even when different parameters or settings are used to generate the stratifications (distance metric, clustering algorithm, and the number of imputations).

The Supplementary Methods include a glossary of technical terms with explanations to elucidate technical terminology.

Input data

ClustALL accepts binary, categorical, and numerical clinical variables as input (e.g., biochemical markers, demographics, clinical scores). Categorical features are transformed internally using a one-hot encoder method, avoiding the assumption of ordinal relationships between categories, which is essential for many clustering algorithms to operate efficiently. A minimum of two features is required, but including more features would lead to more precise clustering. It is important to note that increasing the number of features may also increase the computation time.

Step 1. Data complexity reduction

In this step, highly correlated features are replaced by a reduced set of variables that account for their variability. To that end:

Step 1.1. Dendrogram Hierarchical clustering is applied to the complete dataset, resulting in a dendrogram where variables are grouped based on similarity or collinearity [29]. The depth of each branch represents the distance between the groups of variables. All the possible depths of the dendrogram are extracted, and the sets of variables beneath each depth are stored as *Depth* (see *Glossary of Technical Terms, Supplementary Methods).*

Step 1.2. Preprocessing Principal Component Analysis (PCA) is computed for each set of variables corresponding to each *Depth*, and the first three principal components are stored in a new matrix (*Embedding*) (see *Glossary of Technical Terms, Supplementary Methods)* [30]. For sets that contain only one variable, the variable itself is stored to generate the replacement matrix. This results in a complexity-reduced data set (*Embedding*) for each considered *Depth*. A subset of depths can be considered when the number of variables is too large.

Step 2. Stratification process

In this step, ClustALL calculates and pre-evaluates stratifications for each *Embedding*. For each *Embedding*, the dissimilarity between patients' pairs is computed using correlation-based distance and Gower dissimilarity metric, resulting in two distance matrices [31, 32]. Clustering algorithms are then applied [33–35] depending on the distance used: k-means and hierarchical clustering for correlation distance matrices, and k-medoids and hierarchical clustering for the Gower distance matrix. Throughout all experiments, five different cluster numbers are evaluated $k? \{2, 3, 4, 5, 6\}$. The optimal number of clusters for each strategy is determined based on the consensus from three different measures of clustering internal validation: the sum-of-squares based index or WB-ratio, the Dunn index, and the average silhouette width [36, 37]. The objective is to group patients with comparable data while ensuring that patients in separate clusters are as dissimilar as possible from those in other clusters. As the output for this step, a stratification is derived for each combination denoted as "*embedding+distance metric+clustering method*".

Step 3. Consensus-based stratifications

Step 3.1. Population-based robustness A data-driven threshold is used to define population-based robust subgroups or clusters. For each resulting stratification from the previous step, cluster-wise stability is computed by bootstrapping the dataset 1,000 times and calculating the Jaccard similarity index to the originally defined clusters (see *Glossary of Technical Terms, Supplementary Methods)* [38, 39]. Based on data distribution, stratifications with less than 85% stability (Fig. S4) are excluded. The remaining stratifications are denoted as $Strat_{\text{filt}}$.

Step 3.2. Jaccard distance is applied to compute distances between the population-based robust stratifications [38]. Then, to identify parameter-based robust clusters (where a minor modification in parameter selection provides a similar result), ClustALL considers those combinations that are part of a group of stratifications (green squares in Consensus-based stratifications step in Fig. 1). Then, as initial criteria, that can be modified by the user, centroids from each "combination group" are selected as parameter-based robust stratifications (coloured green squares in Consensus-based stratifications step in Fig. 1). The outcome can be none, one, or multiple ways to stratify the population robustly. In the current analysis, we considered parameter-based robust representatives: centroids of a combination group that includes at least 5 population-based robust stratifications.

ClustALL enables input data with missing values

ClustALL can be adapted to work with missing data (Fig. S1). To that end, the ClustALL method is modified as follows:

Step 1 Adaptation

First, a dendrogram and its associated depths are computed considering the original dataset with missing values. The original dataset is then imputed 1,000 times using the Multivariate Imputation by Chained Equations (MICE) algorithm [40]. MICE is chosen for its capability to manage complex data structures and capture relationships between variables more effectively than other imputation methods [40]. Moreover, MICE offers the flexibility to specify relationships between variables through the *predictorMatrix* parameter within the *mice* function. This parameter allows us to handle interdependencies among input variables by specifying predictors for each target feature, thereby facilitating robust imputation.

Additionally, we employ the most suitable imputation method included in the *mice* function based on the data type. Specifically, predictive mean matching (*pmm*) is utilized for numeric variables, logistic regression (*logreg*) for factor variables with two levels, multinomial logit model (*polyreg*) for factor variables with more than two levels, and ordered logit model (*polr*) for ordered variables with multiple levels. Note that, by implementing imputation iteratively, we mitigate bias by capturing the inherent uncertainty and variability in the process. Subsequently, for each previously calculated *Depth* and each imputed dataset, the Data Complexity Reduction step is applied.

Step 2 Adaptation

Step 2.1 is computed for each combination of depth, distance metric, clustering algorithm and each *Embedding* derived from an imputed dataset. The selection of the optimal number of clusters is based on the consensus from cluster internal validation and the mode of the imputed datasets for each corresponding embedding. Afterwards, a distance matrix (D_{mat}) between individuals is obtained by computing how often two individuals are assigned to the same cluster in each imputation (Fig. S1). Then, D_{mat} calculates a final stratification score using correlation-based distance and hierarchical clustering. In our experience, limited optimization is required here because summarizing the stratification over all imputations separately strengthens what is observed in each imputed dataset. Extra care will be required only in cases where imputations may differ significantly. After this modification, the method follows as previously described (Fig. S1).

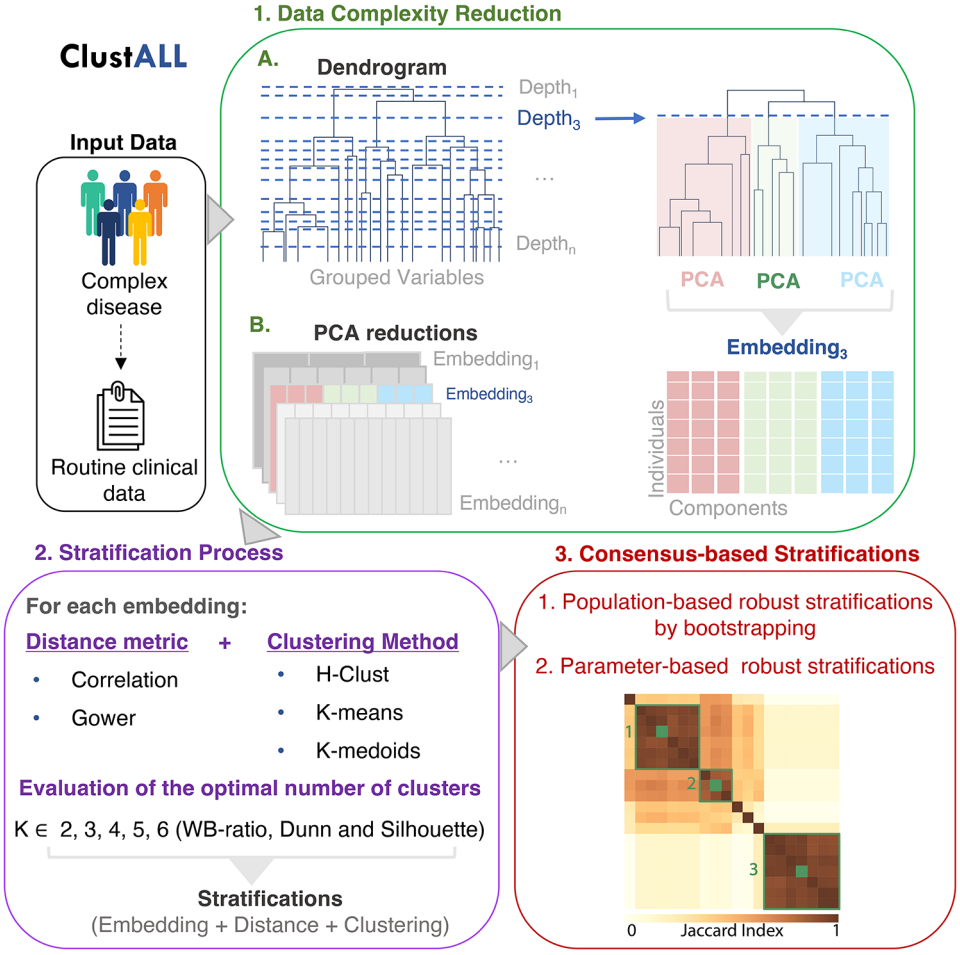


Fig. 1 Schematic overview of the different steps of ClustALL approach (best viewed in colour). ClustALL takes clinical variables as input. First, data complexity is reduced by grouping the features into a dendrogram, assessing the resulting depths, and using Principal Component Analysis (PCA) (green panel). The output is an embedding for each possible depth. Then, stratification is computed considering the combination of different distance measures, clustering techniques, and cluster numbers (K) (purple panel). In the final step, non-robust stratifications are filtered, and the centroids derived from computing Jaccard (coloured green squares) similarity among the robust stratifications (green squares) are considered the final representatives of the stratifications (red panel)

Data source

The data utilized in this study were obtained from two independent multicentre studies: the European PREDICT cohort and the Latin-American ACLARA cohort, conducted as part of the European Project DECISION [28, 41]. Both cohorts collected various measures, including clinical, pharmacological, biomarker, and outcome data from patients with acute decompensation of cirrhosis upon hospital admission and during follow-up visits. The follow-up period was 90 days for the PREDICT cohort and 28 days for the ACLARA cohort. To be eligible for the present study, patients were required to have acute decompensation of cirrhosis upon hospital admission, with available information on short-term outcomes, drug intake, and available biological samples. Ultimately, 766 patients from the PREDICT cohort and 580 patients from the ACLARA cohort and 74 features (continuous and categorical) were included in the analysis. The features included demographic information, clinical and laboratory data, medical history, risk factors, and cirrhosis scores at hospital admission, with missing values accounting for less than 30% (Table 1). To avoid bias from missing data, imputation was performed with 1,000 iterations using the MICE method [40].

ClustALL comparison to different clustering methodologies

A comparison was conducted between the ClustALL framework and classical clustering algorithms. Stratification was performed on 1,000 imputed datasets using classical k-means and hierarchical methodologies with k values of 2 and 3, considering that ClustALL robust stratifications comprised two or three patient subgroups. Bootstrapping was performed for the classical clusters to evaluate cluster-wise stability [42]. The resulting stability was compared to ClustALL stability through the Kolmogorov-Smirnov test. Moreover, the clinical utility

Table 1 Complete list of input features. Patient characteristics included in the analysis: demographics, cause of cirrhosis, main reason for hospitalization, manifestations at admission, cirrhosis severity scores, medical history, lifestyle and laboratory variables

Demographics	Age, sex, height[+], weight[+], BMI, ethnicity (Black or African American, Asian, White, other)
Cause of cirrhosis	Alcohol, viral, alcohol + viral, NASH, cryptogenic, other
Main reason for hospitalization	Ascites, hepatic encephalopathy, gastrointestinal bleeding, spontaneous bacterial peritonitis, other infection
Manifestations at admission	Clinical Events (ascites, hepatic encephalopathy, gastrointestinal bleeding, acute kidney injury, bacterial infection, acute alcoholic-steatohepatitis, acute viral hepatitis, hepatocellular carcinoma), number of clinical events (the sum of clinical events), number of precipitating events (the sum of precipitating events: proven bacterial infection, acute alcoholic-steatohepatitis, CLIF-C AD > 50), organ dysfunction (liver, renal, cerebral, coagulation, cardiac, respiratory), number of organ dysfunctions (the sum of organ dysfunctions), organ failure (liver, cerebral, coagulation, cardiac, respiratory), number of organ failures (the sum of organ failures)
Cirrhosis Severity Scores	Child-Pugh, CLIF-C AD, CLIF-C OF, MELD, MELDNA
Medical history	History of diabetes, history of hypertension, history of previous decompensations
Lifestyle	Alcohol, active alcohol consumption[+], tobacco
Laboratory variables (measured in serum)	Alanine transaminase, aspartate aminotransferase, albumin, bilirubin (total), gamma-GT, C-reactive protein, sodium, potassium, glucose[+], hemoglobin, hematocrit, creatinine, white blood cell count, lymphocytes, monocytes, neutrophils, INR (International Normalized Ratio), platelet, SpO2 (%), SpO2/FiO2 Ratio

[+]Variables not included in ACLARA cohort

of the various stratifications was assessed by examining the clinical insights obtained from the different clusters.

Statistical methods

All analyses were performed in the R Computing Environment version 4.0.3 [43].

Descriptive statistics

Descriptive characteristics of the PREDICT and ACLARA study populations were reported as means with standard deviations for continuous variables and proportions of patients for categorical variables.

Feature analysis

The identification of the minimal-size predictive signatures with maximal predictive power leading to each stratification was performed using the *fbed.reg* function with default hyperparameters from the MXM R package [44, 45].

Parametric tests

Differences between clusters in the PREDICT and the ACLARA cohorts were assessed using one-way ANOVA for continuous variables, while binary variables were tested with the chi-square test. The association between the PREDICT clusters identified with ClustALL –exclusively using data obtained at admission– with the groups of patients based on their clinical course [28], was tested with the Fisher test.

Stratification model reproducibility

AD-strat model was validated in a separate cohort of patients with acute decompensation of cirrhosis from the ACLARA cohort and in the PREDICT follow-up time points. For this purpose, the k-nearest neighbours (kNN) model was trained on the PREDICT *AD-strat* cluster labels based on the signatures previously defined as most predictive in the feature analysis [46]. The K parameter was selected based on different measures that assessed the overall model performance over different K's [47], including accuracy, the area under the curve (AUC), error rate (ER), false positives (FP), and false negatives (FN) (Table S7). After applying the kNN algorithm, the target data (ACLARA cohort and PREDICT follow-up) was labelled based on the majority votes from the kNN and imputed datasets. Deeper details on the kNN model and its performance evaluation are included in the Glossary in the Supplementary Methods.

Survival analysis

Cumulative incidences of ACLF development and liver-related death were estimated using the cumulative incidence function of the survival R Package. Liver transplantation was considered a competing event. A p-value lower than 0.05 with Benjamini and Hochberg (BH) adjustment was considered statistically significant.

Longitudinal analysis and model evaluation

All PREDICT patients with ≥1 post-baseline assessment (n=688) were included in longitudinal outcomes analyses for a period of 90 days after hospital admission. Sankey diagrams were generated to show the patients' transfers among the *AD-strat* clustering, liver transplant, ACLF development, death, and survival status. The predictive power of the stratification models at follow-up time points versus at baseline in the PREDICT cohort was evaluated using the Bayesian Information Criterion (BIC), the Akaike information criterion (AIC), the concordance, and the Likelihood ratio goodness-of-fit parameters [48, 49].

Results

ClustALL, a robust data-driven framework for patient stratification in complex diseases

We developed a specialized stratification framework, referred to as ClustALL, specifically designed to accurately identify all potential alternatives for stratifying a population using clinical multimodal data at hospital admission as input. The ClustALL methodology consists of three main steps illustrated in Fig. 1 and detailed in the Methods section: 1) **Data Complexity Reduction** (depicted in the Green Panel of Fig. 1) aims to simplify the original dataset by mitigating the impact of redundant information (highly correlated variables). As a result, we obtain a set of embeddings, each one derived from different groupings of clinical variables. 2) **Stratification Process** (depicted in the Purple Panel of Fig. 1), where, for each embedding, multiple stratification analyses are performed using different combinations of among the most widely used distance metrics and clustering methodologies (REF). From each combination, denoted as "*embedding+distance metric+clustering method*", a stratification is derived. 3) **Consensus-based Stratifications** step (depicted in the Red Panel of Fig. 1) aims to identify robust stratifications that, in addition, exhibit minimal variation when combination parameters ("*embedding+distance metric+clustering method*") are *slightly* modified. ClustALL performs a **population-based robustness** analysis for each stratification using bootstrapping. This analysis ensures that combinations associated with non-robust stratifications are excluded. The resulting stratifications are then compared using the Jaccard distance. As a result, a heatmap is generated to visually identify groups of representative stratifications (green squared lines). The selection of representative stratifications enables the preservation of those stratifications that demonstrate **parameter-based robustness**: consistency even when various parameters, like distance metrics or clustering methods, are altered. For each group of stratifications, the centroid is selected as the final stratification (green squares).

Combining **these three steps allows ClustALL to identify none, one, or multiple robust stratifications in a given population** of patients with complex diseases. Importantly, a specific implementation of ClustALL is designed to **effectively handle datasets with missing data effectively**, ensuring that incomplete information does not hinder the stratification process.

ClustALL uncovers stratification in a cohort of patients with acutely decompensated cirrhosis: a proof-of-concept

Study population

The ClustALL approach was applied to a subset of individuals from the European PREDICT cohort [28], which included 766 patients with acute decompensation of cirrhosis and 74 clinical features collected at hospital admission, with less than 30% missing values. Complete information on patient characteristics and short-term outcomes, including acute-on-chronic liver failure (ACLF), liver transplant, and death, can be found in Supplemental Table 1.

ClustALL identified five different alternatives to stratify the population

The ClustALL workflow was utilized to discover potential new sub-phenotypes of patients with acute decompensation of cirrhosis within the PREDICT cohort upon hospital admission (Fig. 2). To handle missing values in the dataset, we employed the ClustALL framework, which incorporates imputations using 1,000 iterations, as described in the Methods section. The Data Complexity Reduction Step resulted in 72,000 embeddings (Fig. 2). The Stratification Process generated 288 stratifications based on the different combinations of "*embedding+distance metric+clustering method*" (Fig. 2). Among these, 144 population-based robust stratifications were identified through the Consensus-based Stratifications step, resulting in five groups of parameter-based representative stratifications. The centroid was selected for each group of stratifications (Fig. 2).

ClustALL provides better resolution than classical clustering tools

We conducted an analysis to assess the added value of ClustALL when compared with classical clustering methodologies such as k-means or hierarchical clustering. Regarding the classical methodologies, our findings revealed that when using correlation as a distance metric, 90% of patients were consistently assigned to a single cluster, regardless of the number of clusters considered; when Gower distance was utilized, the distribution of patients across clusters presented a more balanced distribution (Table S2). Notably, the population-based robustness of the stratifications generated by ClustALL was significantly higher (p-value <0.01) compared to the results obtained using k-means and hierarchical clustering (Fig. S3). In summary, our observations demonstrate that ClustALL significantly outperforms classical methodologies regarding population-based robustness.

Characterization of the five robust stratifications within the PREDICT population

After identifying the robust stratifications, we aimed to explore and characterize the distinct clusters observed in each of the five alternative stratifications. These stratifications divided the patients into two clusters, except for stratification 1, which had three clusters. We visually investigated the separation by representing each stratification in a low-dimensional space using the

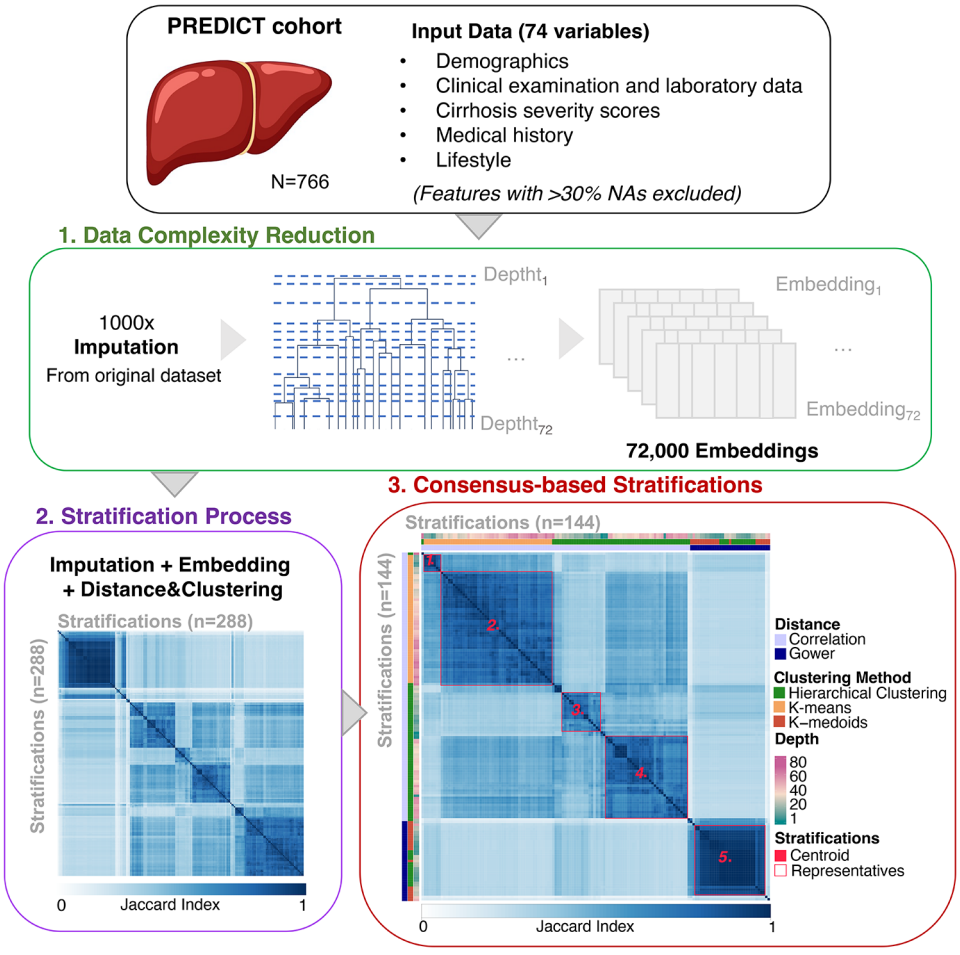
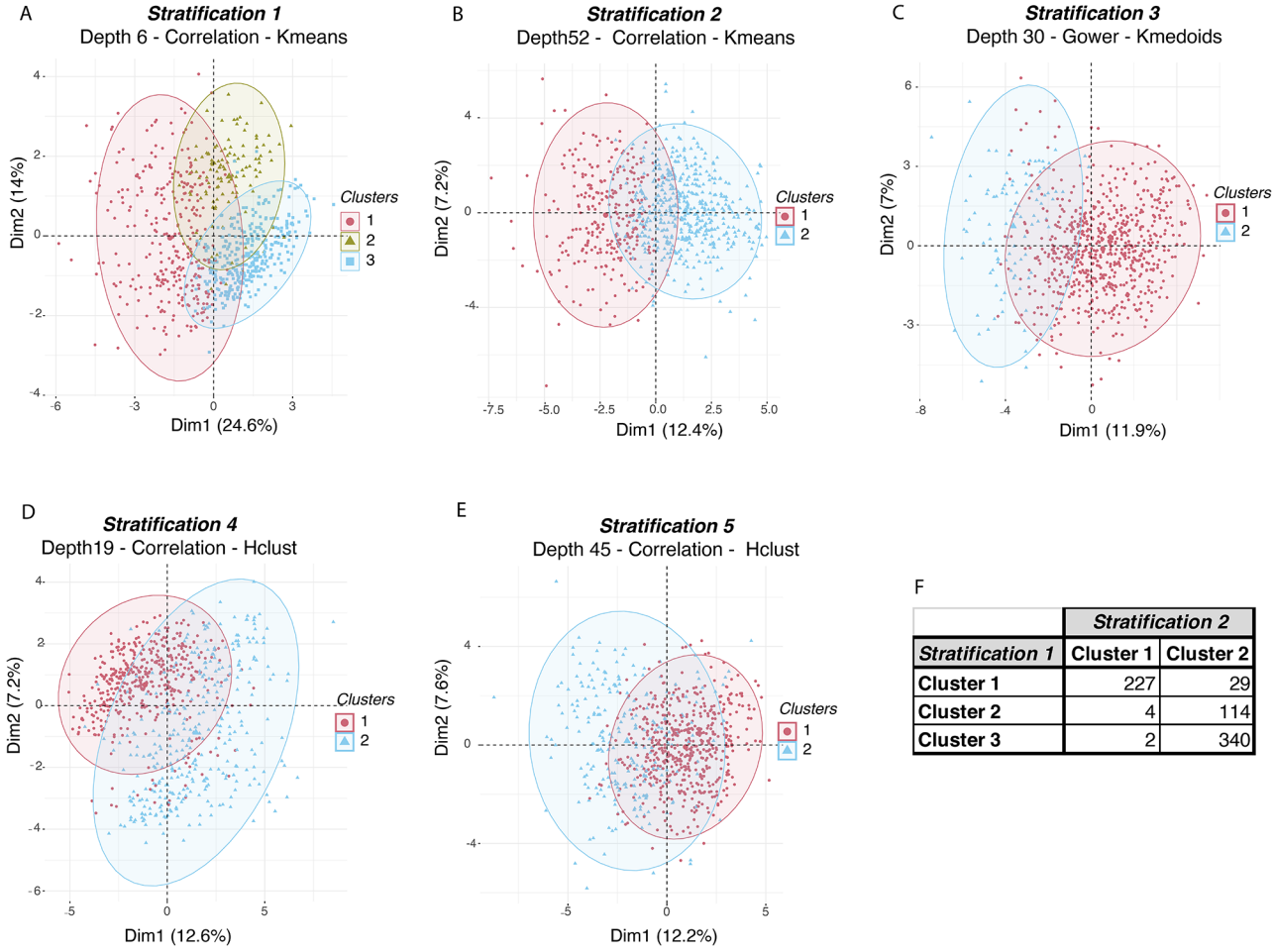


Fig. 2 Summary of the outputs from the different steps of the ClustALL framework when applied to the PREDICT cohort (*N* = 766). Input data comprised 74 clinical features with less than 30% missing values. The analysis utilized 1,000 imputed datasets. The Data Complexity Reduction step (green) was applied to 72 depths of the 1,000 imputed datasets. The Stratification Process step (purple) considered various clustering combinations resulting in 288 stratifications. After bootstrapping, 144 robust stratifications remained. Finally, in the Consensus-based Stratification step (red), five groups of robust stratifications (red squares) were identified, and the centroid was selected from each group as the final stratifications (red coloured squares)

corresponding embeddings derived from the dendrogram depths (Fig. 3A-E) and the complete dataset (Fig. S2). Further exploration revealed that stratification 1 was a subdivision of stratification 2 (Fig. 3F). We then determined the minimal sets of variables (excluding the cirrhosis severity scores (Table S1 variables 44 to 48)) with the highest predictive performance in differentiating the clusters for each stratification (Tables S3–S7) [44, 45]. The different classification approaches were described by 25 variables from a total of 74 (Table S1 variables 1 to 74), with 8 to 12 variables per stratification (Fig. 4A). Notably, all stratifications included 3 common features: (i) serum bilirubin concentration (either as a continuous variable or categorized under the term "liver dysfunction" [50]); (ii) INR (either as a continuous variable or categorized under the term "coagulation dysfunction" [50]; (iii) the number of organ dysfunction or failure. Precipitating events were present in all but one stratification (stratification 3) either as a sum or individually (gastrointestinal bleeding, alcohol-related hepatitis, acute viral hepatitis). Diabetes mellitus was included in two stratifications. Conversely, age, sex, BMI, cause of cirrhosis, and lifestyle were present in no or one stratification. Interestingly, stratification 1 and 2 shared almost the same minimal set of variables. Both stratifications identified a group of patients with a severe phenotype attested by low serum sodium, low serum albumin, high serum bilirubin, high INR, high C-Reactive Protein (CRP) and leucocytes, and the number of precipitating events (Fig. 4B). Hepatic encephalopathy was present in stratification 1 but not in 2 [51]. A complete statistical characterization of the stratifications is provided in Tables S3 to S7. Considering the clinical implications of the features and the finer classification of the patients, we identified stratification 1 as the most insightful for further exploration in patients with acute decompensation of cirrhosis. Henceforth, in our discussions, we will refer to this specific stratification as *'AD-strat'*.



F

	Stratification 2	
Stratification 1	Cluster 1	Cluster 2
Cluster 1	227	29
Cluster 2	4	114
Cluster 3	2	340

Fig. 3 Principal Component projection of the ClustALL robust stratifications based on the embedding associated with each stratification. (**A**-**E**). Low-dimension representation of the robust stratifications after applying the ClustALL framework to the PREDICT cohort. For each one of the 5 robust stratifications identified by ClustALL, the Principal Component Analysis of the Embeddings corresponding to the specific dendrogram depth associated with the stratification is shown. The *x* (Dim1) and *y* (Dim2) axes represent the first and second principal components respectively, which are linear combinations of the original variables. (**F**). The overlap between the clusters in stratifications 1 and 2 shows that stratification 1 is a subdivision of stratification 2

AD-strat provides prognosis value

The *AD-strat* stratification is defined by three subgroups (clusters) of patients with acutely decompensated cirrhosis, revealing different clinical characteristics and disease progression. Cluster 1 included 306 patients (39.95%) who exhibited the most clinically critical scenario (Fig. 5A, B and Table S3). These individuals had the highest rates of organ dysfunction, clinical events, and precipitating events (Table S3). They had a marked acute inflammatory profile (high white blood cell count and CRP level), poor liver function (low levels of albumin and high levels of INR and serum bilirubin), and more hepatocyte injury (higher levels of serum aspartate aminotransferase). Conversely, Cluster 2 (n=118; 15.4%) and Cluster 3 (n=342; 44.6%) had a less severe presentation. The main difference between Cluster 2 and 3 was hepatic encephalopathy, found in 89% of the patients in Cluster 2 and almost no patients in Cluster 3 (Fig. 5A, B and Table S3). Importantly, a significant prognostic value of *AD-strat* was revealed by exploring the cumulative incidence of ACLF and death over 90-day follow-up (Fig. 5C).

Liver transplantation was considered a competing event as it represents a definitive intervention that dramatically changes the course of the disease, offering a potential cure for end-stage liver disease, similarly as in other studies [28, 52, 53]. Patients in Cluster 1 had poor short-term outcomes, with a cumulative incidence of ACLF and death, both by 90 days of 24.1 and 21.5, respectively. While Clusters 2 and 3 had similar risks of ACLF by 90 days (8.6% and 10.2%, respecively), the risk of death by 90 days was lower for Cluster 2 than Cluster 3 (4.3% vs. 10.7%). When we compared the clusters identified with ClustALL - exclusively using data obtained at admission - with the groups of patients based on their clinical course [28], we found a statistically significant association (Fisher test, p-value<0.01) (see Table S8).

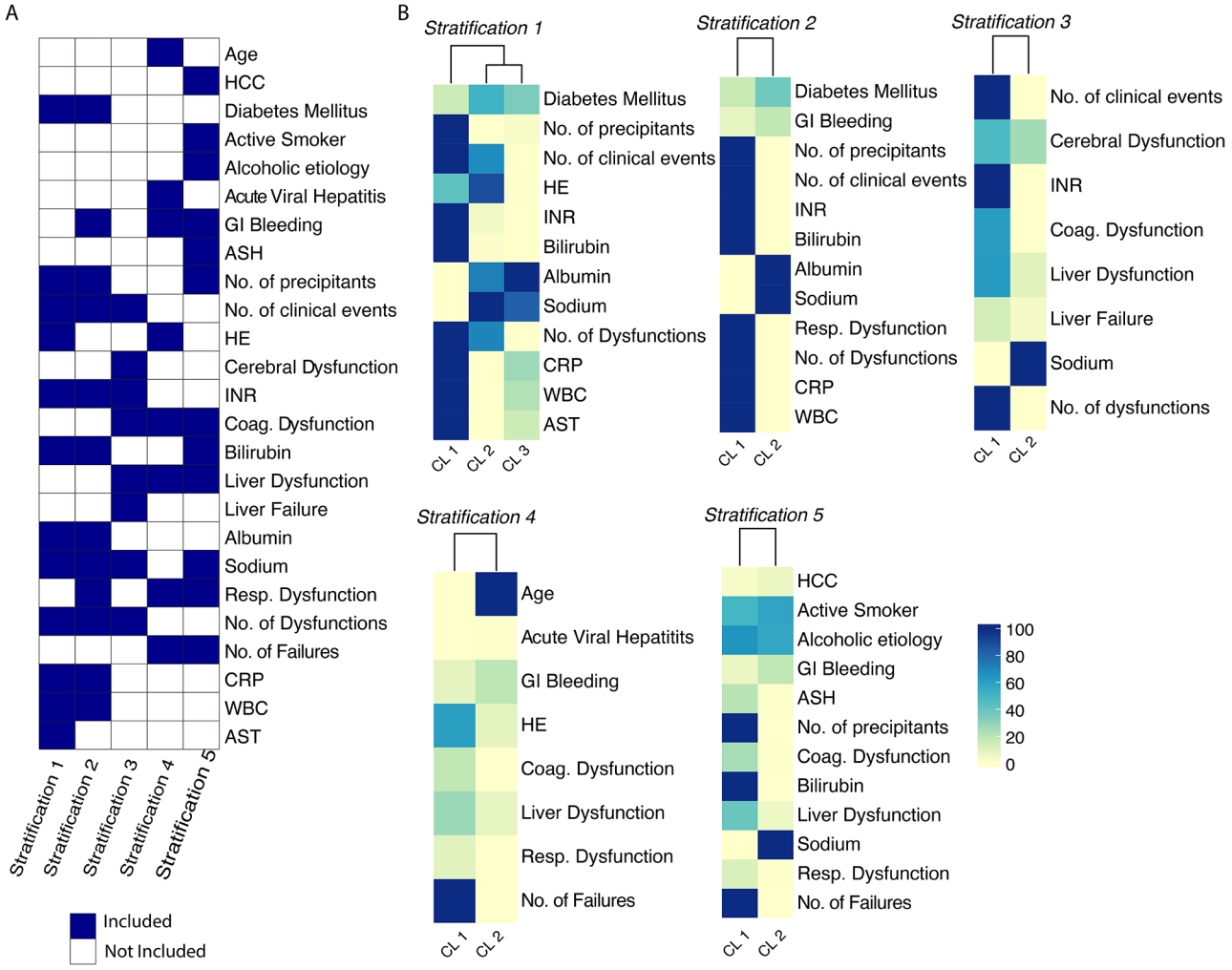


Fig. 4 Overview of the variables driving the ClustALL stratifications. (**A**). Heatmap with the minimal set of variables required to describe the 5 different stratifications, accounting for 25 out of 74 input variables. (**B**). Heatmaps of the minimal set of patient characteristics per stratification. The heatmap colour scale depends on the data type. In the case of binary variables, the value indicates the percentage of patients with such binary characteristics, e.g., the presence of Diabetes Mellitus. For continuous variables, the colour scale represents a scaled value from the highest cluster mean (100.0) to the lowest cluster mean (0.0), e.g., Albumin and CRP. Abbreviations: ASH = Acute Alcoholic-Steatohepatitis, AST = Aspartate aminotransferase, CL = Cluster, CRP = C-Reactive Protein, HE = Hepatic encephalopathy, HCC = Hepatocellular Carcinoma, INR = International normalized ratio, WBC = White blood cell counts

We observed that 61% of patients with pre-ACLF were in Cluster 1, and 48% of patients with stable decompensated were in Cluster 3.

Reproducibility of the stratification model in an independent cohort

We assessed the validity of the *AD-strat* model in a large independent prospective multicentre cohort that included 580 patients with acute decompensation of cirrhosis from the Latin-American ACLARA study [41]. Using as a reference the PREDICT *AD-strat* clusters, we labelled ACLARA patients using the k-nearest neighbours classification algorithm (Table S10) [46]. The classification model included the 12 predictive variables previously identified in the feature importance analysis (Fig. 4B Stratification 1). Importantly, the allocation of the patients to the clusters was consistent and independent of the imputation in 99% of the cases (Fig. 6A), and the distribution of individuals by *AD-strat* clusters within ACLARA closely mirrored that of the PREDICT cohort (Fig. 6B). As expected, the clustering of the ACLARA cohort exhibited similar clinical feature patterns to the PREDICT cohort (Figs. 4B and 6C Stratification 1). Furthermore, the features describing the subgroups demonstrated statistical significance (Table S11). Finally, we assessed the clinical relevance of the clustering in terms of prognosis, specifically examining the short-term outcomes available in the ACLARA cohort 28 days after hospital admission. Similar to results obtained in the PREDICT cohort, Cluster 1 displayed a bad prognosis for both ACLF and death, while Cluster 3 showed a better prognosis (Fig. 6D). In ACLARA, all patients from

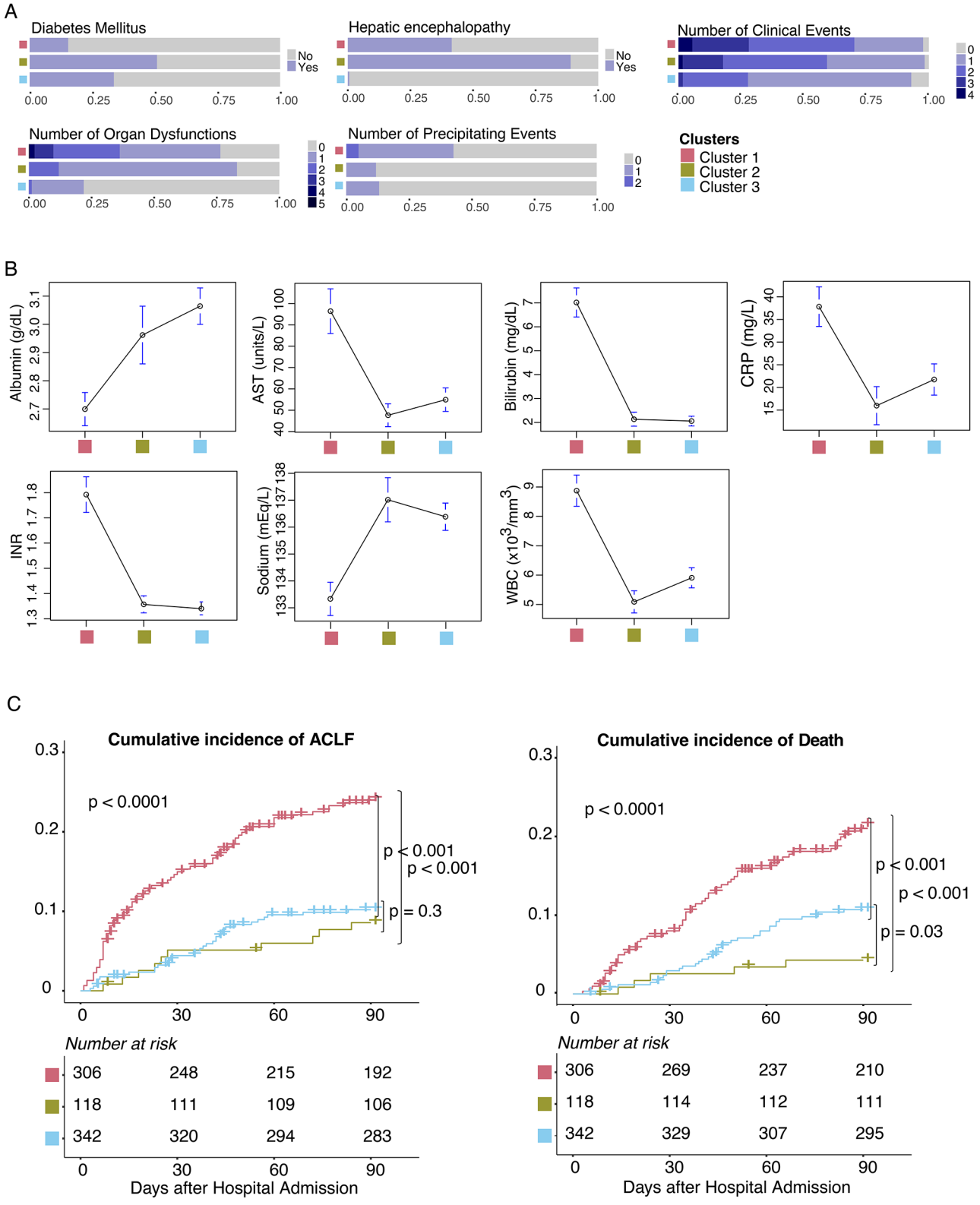


Fig. 5 Clinical overview of the AD-strat derived clusters in the PREDICT cohort. (A, B). Distribution of the highest predictive performance-related patient characteristics among AD-strat clusters; (**A**) categorical variables, (**B**) numerical variables. (**C**) Cumulative incidence of ACLF (left) and death (right) according to the AD-strat clustering in PREDICT cohort considering 90 days after hospital admission, with the number of patients at risk per cluster (Transplantation counted as a competing risk to death). Abbreviations: AST = Aspartate aminotransferase, CRP = C- Reactive Protein, INR = International normalized ratio, WBC = White blood cell counts

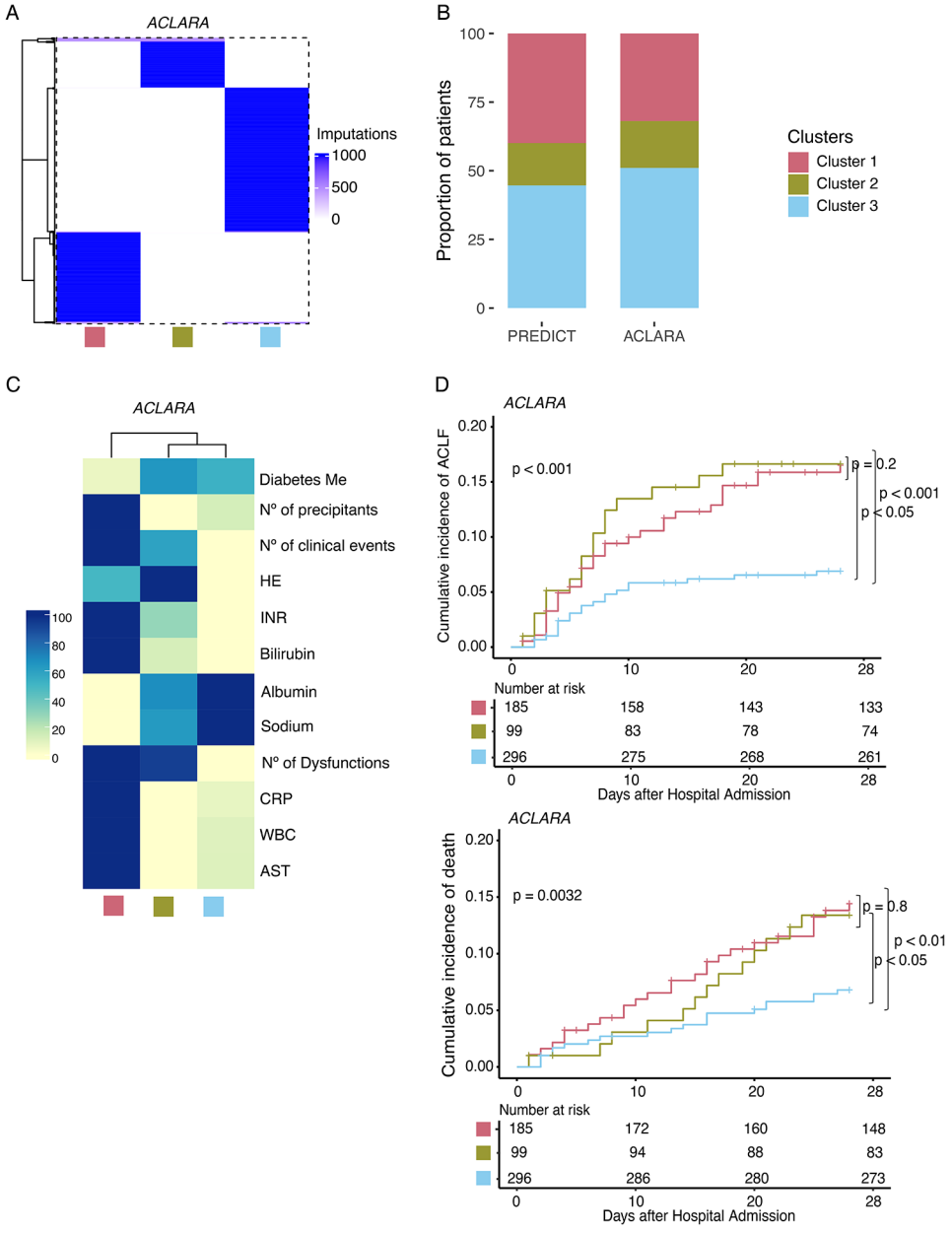


Fig. 6 Reproducibility of the AD-strat model in the ACLARA cohort. (**A**) Distribution of the labels in the ACLARA cohort after applying the kNN model 1,000 times. (**B**) Proportion of patients distributed in the 3 clusters in the PREDICT and the ACLARA cohorts. (**C**) Heatmap of patient characteristics per cluster in the ACLARA cohort. Bars on the right show the colour scale representing the proportion with each binary characteristic, such as diabetes. Continuous variables, such as bilirubin, represent a scaled value from the highest cluster mean (1.0) to the lowest cluster mean (0.0). (**D**) Cumulative incidence of ACLF (up) and death (down) according to the AD-strat clustering in ACLARA cohort considering 28 days after hospital admission, with the number of patients at risk per cluster (Transplantation counted as a competing risk to death). Abbreviations: AST = Aspartate aminotransferase, CRP = C- Reactive Protein, INR = International normalized ratio, WBC = White blood cell counts

Cluster 2 were afflicted by hepatic encephalopathy (Table S11) and showed a poor prognosis similar to that of Cluster 1. Ethnicity was homogeneously distributed across clusters (Table S2). In particular, Native Americans represented 21% of Cluster 1, 15% of Cluster 2, and 14% of Cluster 3. Complete information on patient characteristics and short-term outcomes is reported in Supplemental Table 9.

AD-strat as a marker for clinical management

Finally, we investigated the clinical value of the stratification during the follow-up visits of the PREDICT cohort. Based on the PREDICT study design [28], two follow-up visit plans were established depending on the reported disease severity (CLIF-C AD-score) at hospital admission (Fig. 7A). For patients with a CLIF-C AD-score ≥ 50, the scheduled visits were performed at hospital admission

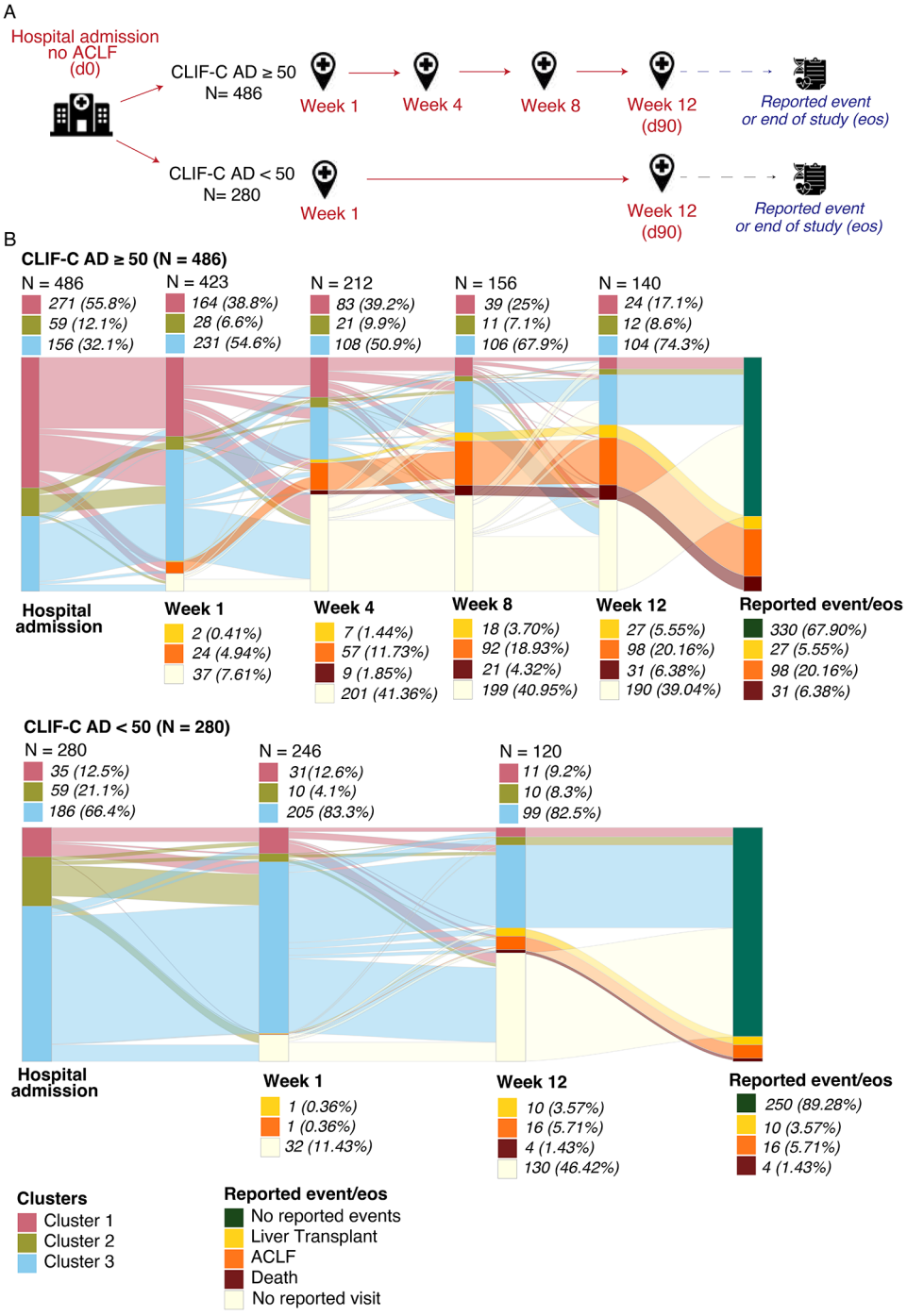


Fig. 7 Distribution and transition of the AD-strat derived clusters at different visits in the PREDICT cohort. (**A**). Schematic representation of PREDICT study design. Two follow-up visit plans were defined according to the reported disease severity (CLIF-C AD-score) at hospital admission (red). The information about the occurrence of any adverse event (liver transplant, ACLF or death) during the whole visit plan or the absence of events at the end of the study was tracked (blue). (**B**) Sankey plots show the cluster label of each patient over the follow-up visits. The follow-up flows of patients with CLIF-C AD >= 50 at hospital admission (up) and CLIF-C AD < 50 at hospital admission (down) are shown. The distribution of the patients assessed at each follow-up visit per cluster is shown as frequency and proportion on the top of the Sankey representations. The accumulated frequency and proportion of adverse events at each follow-up visit respecting the whole cohort (for CLIF-C AD >= 50, $n = 486$; for CLIF-C AD < 50, $n = 280$) are shown on the bottom of the Sankey representations. Reported event/end of study (EOS), shows the status of a patient at the "end of the study": patients with a reported event or patients with no reported event

and 1, 4, 8 and, 12 weeks after enrolment. For patients with a CLIF-C AD-score < 50, the scheduled visits were performed only at hospital admission and 1 and 12 weeks after enrolment.

Of the 766 patients included in the PREDICT study, 688 had at least one follow-up visit. For this subset of patients with available data, we labelled each of them at each follow-up visit using the kNN algorithm (Fig. 7B). This approach allowed an overview of the patient stratification over the entire study duration and revealed the patient flow over time, highlighting cluster transitions.

Consistent with the previous *AD-strat* characterization at hospital admission (Fig. 5 and Table S3), we identified more than 50% of patients with a CLIF-C AD score ≥ 50 (n = 486) were classified as Cluster 1, while patients with CLIF-C AD score < 50 (n = 280) were predominantly classified as Cluster 3 (66.4%) (Fig. 7B). Changes in cluster proportions were observed during the patients' follow-up. Stratification changes over time were more pronounced among patients with a CLIF-C AD scores ≥ 50 at hospital admission, showing a progressive reduction of patients classified as Cluster 1 (55.8% at hospital admission, 38.8% at week 1, 39.2% at week 4, 25% at week 8, and 17.1% at week 12) and an increase of those classified as Cluster 3 (32.1% at hospital admission, 54.6% at week 1, 50.9% at week 4, 67.9% at week 8, and 74.3% at week 12). Additionally, there was a progressive increase in the proportion of patients classified as Cluster 3 for those patients with a CLIF-C AD-score < 50 at hospital inclusion (66.4% at hospital admission, 83.3% at week 1, and 82.5% at week 12).

To assess the effectiveness of the *AD-strat* throughout disease progression, we determined its prognostic value in two scenarios: (1) using the stratification at hospital admission, and (2) using the stratification at the last visit reported before the occurrence of any adverse event (we considered any visit between week 1 and 12) or at the end-of-study (EOS) (week 12 visit). A significant difference was observed (p < 0.001, Wilcoxon test) when comparing the time window between the visit used in each scenario and the occurrence of adverse events (Fig. S5), indicating that in the second scenario, we evaluated patients during a visit much closer to the event.

Ultimately, the cumulative incidence of ACLF and death as stratified at the last visit demonstrated a more significant separation between clusters compared to patient stratification at hospital admission (Fig. 8). There was an increase in the incidence for those patients classified as Cluster 1 (18.46% and 18.45% at baseline and 28.16% and 26.8% at the last visit for ACLF and death, respectively). Accordingly, the goodness-of-fit parameters indicated an improvement in risk prediction with the last visit stratification, suggesting an enhanced predictive power as the event approached (Table S12).

Discussion

In the current era of personalized medicine, there is a growing focus on elucidating the complexities of disease populations, reflecting an emphasis on understanding their inherent heterogeneity [54–56]. Consequently, both academic and clinical efforts have been dedicated to characterizing disease subtypes for the purposes of identification, treatment, and prognosis. Or more general, aiming to enhance our understanding and management of complex conditions. Furthermore, the WHO has recognized patient stratification as an invaluable approach [8]. It is important to note that patient stratification extends beyond mere outcome prediction scores, particularly in scenarios where a "one-size-fits-all" approach to treatment may inadequately address the diverse needs and characteristics of individual patients [57].

Patient stratification, as investigated in this study, involves the unsupervised grouping of patients based on available clinical data. Interestingly, while significant progress has been made in classification problems, particularly in domains like single-cell transcriptomic analysis [58, 59], unsupervised clustering of patients based on clinical information is still in the developmental stage [9, 60]. Notably, the existing challenges in clinical stratification, such as handling mixed data types, missing values, or highly correlated variables, are often mitigated using ad-hoc solutions, given the absence of a comprehensive method to address them. To overcome the aforementioned idiosyncrasies, we have developed a novel computational framework named ClustALL.

During the development of ClustALL, our focus extended beyond simply generating patient groups; we were equally invested in ensuring the robustness of the identified stratifications. Typically, clustering robustness can be evaluated based on the stability of the clusters when modifications are made to the population using methods such as resampling or bootstrapping (population-based robustness). Significantly, ClustALL incorporates a second, less explored but equally important, dimension of robustness: assessing the consistency among the resulting stratifications when minor modifications are applied to the clustering parameter settings. This property has already been explored in the context of gene expression data as the "propensity of a clustering algorithm to maintain output coherence over a range of settings" [61]. Another major feature of ClustALL, is its capacity to identify more than one robust stratification within a given population. Clinical data is complex and allows for multiple uses and "multiple interpretations" that *may* result in several *valid* groupings [62]. Indeed, the concept of "multiple interpretations" arises from how variables are utilized in the clustering process and has been a research subject in the last decades [63]. Traditional methods such as k-means or hierarchical clustering

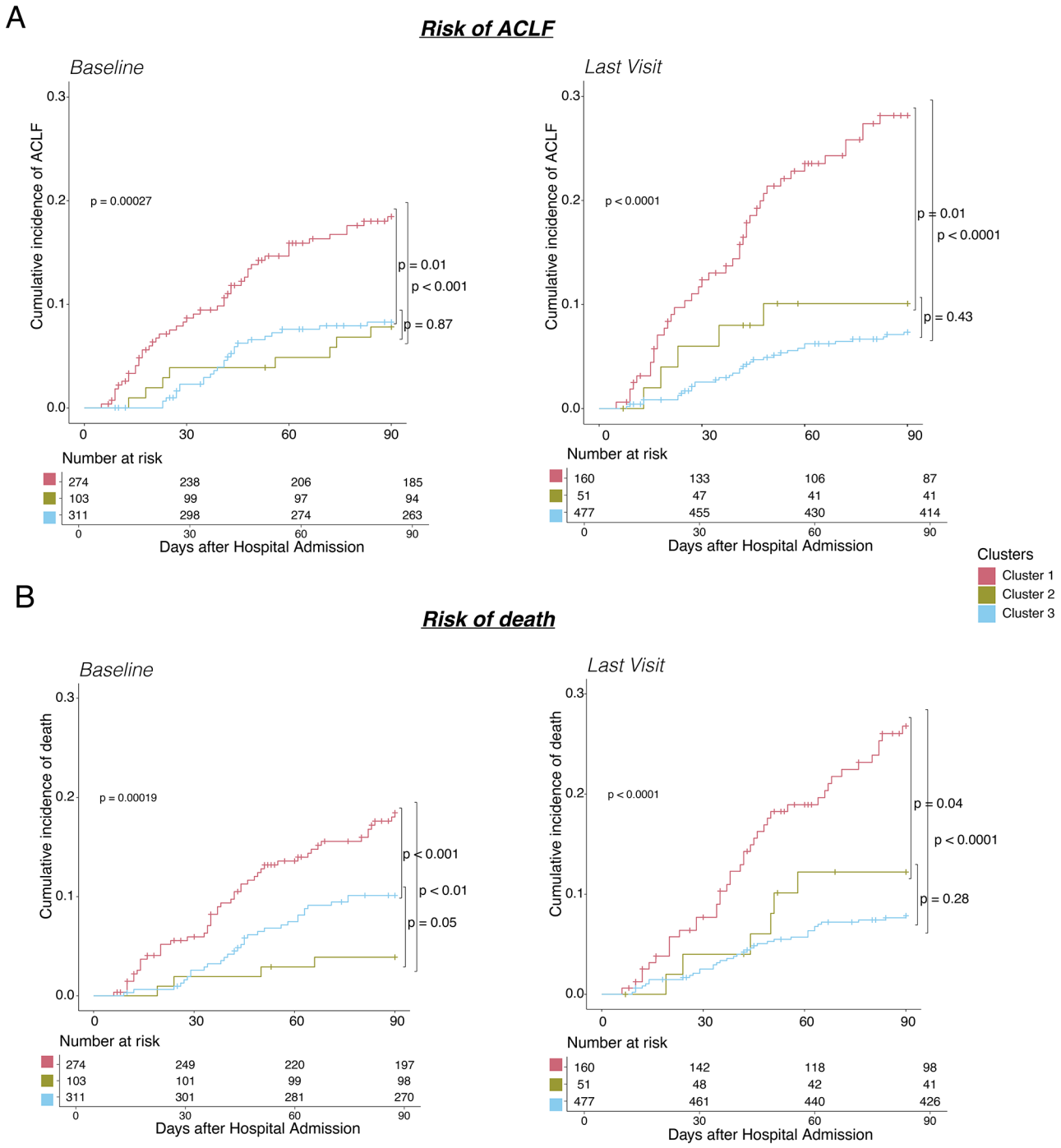


Fig. 8 Assessment of the risk of adverse events according to the AD-strat clusters at different time points. (**A**, **B**). Cumulative incidence of ACLF (**A**) and death (**B**) according to the AD-strat clustering in PREDICT cohort at hospital admission (left) versus at last visit (right) considering 90 days after hospital admission, with the number of patients at risk per cluster (Transplantation counted as a competing risk to death)

typically yield a single outcome, which may be influenced by random initial conditions at the start of the algorithm. We consider that any stratification method should allow for the identification of multiple solutions, necessitating clinical feedback to ascertain their relevance. In contrast to traditional methods, following trace-based clustering principle, ClustALL does not rely on a random single initialization of the clustering, but, in general, integrates the information of multiple clustering efforts and evaluation criteria [18]. In summary and considering all these factors, we believe that ClustALL represents a necessary step towards practical unsupervised patient stratification;

notably through the incorporation of parameter-based robustness and its capacity to identify more than one stratification.

To assess the effectiveness of ClustALL, we applied it as a proof-of-concept in a cohort of patients with acutely decompensated cirrhosis, considering clinical data collected at hospital admission. Such an attempt to apply a data-driven stratification to patients with cirrhosis has never been conducted. The stratification we set up differs from the scores developed and routinely used in patients with cirrhosis (e.g., MELD, MELD-Na, Child-Pugh, CLIF-C-AD) both in terms of design and use. Indeed, all these scores were built using a follow-up endpoint (usually death) in patients receiving therapies. These scores are helpful in identifying patients at high risk of poor outcomes. Still, they do not fully capture the heterogeneity of the patients at admission for several reasons: (a) some features explaining patients' heterogeneity might not have an independent prognostic value, either because the prognostic information they carry is contained in other variables, or because therapies administered to patients during their follow-up blunt their impact; (b) a similar survival rate does not imply similar pathophysiological mechanisms. For instance, in PREDICT, clusters 2 and 3 have a similar rate of ACLF, while they strongly differ with regard to the prevalence of hepatic encephalopathy.

In the first step of our analysis, we identified five alternative stratifications for patients with acute decompensation of cirrhosis. Interestingly, all stratifications included markers of impaired liver function, serum bilirubin and INR, and the number of organ dysfunction or failure, and all but one included precipitating events. This emphasizes that these features are crucial when designing a clinical trial, including patients with acute decompensation of cirrhosis. Our data-driven approaches show that serum bilirubin and INR are not only key to predicting the outcome of patients with cirrhosis but also to explaining heterogeneity at admission. On the contrary, features like age, sex, BMI, cause of cirrhosis, and lifestyle were present in no or only one stratification, suggesting that these features are not key when designing a clinical trial. The stratification we selected (*AD-strat*) provided a more granular resolution by allowing the identification of three subgroups of patients.

In this stratification, diabetes mellitus is taken into account. While it is known that diabetes is an independent risk factor for cirrhosis decompensation [64, 65], the role of diabetes once acute decompensation has happened has been overlooked so far. This place of diabetes is quite unique since causes of cirrhosis, comorbidities, or lifestyle were not part of the key features of *AD-strat*. Hepatic encephalopathy strongly impacted the categorization of patients with acutely decompensated cirrhosis. Notably, 89% and 100% of the patients in Cluster 2 from the PREDICT and ACLARA cohorts, respectively, presented hepatic encephalopathy at the time of hospital admission. This may explain the intermediate prognosis observed in patients within Cluster 2, as hepatic encephalopathy is recognised by its fluctuating nature and potential reversibility [66, 67]. The dynamic nature of hepatic encephalopathy may also explain why Cluster 2 was not a static group over time [68].

The stratification presented here is not intended to guide clinical bedside decisions or to create a new prognostic score, but rather to identify *more* homogeneous patient populations upon hospital admission. However, once we applied ClustALL, we observed that the three subgroups of patients identified had a different outcome. Moreover, employing *AD-strat* labelling over time facilitated dynamic and enhanced identification of high-risk patients in the PREDICT cohort. These findings underscore ClustALL's ability not only to stratify patients based on baseline characteristics, with a prognostic relevance. In this regard, *AD-strat* might be a useful tool for designing future clinical trials by including more homogeneous patient populations. Using ClustALL may also offer insights into applying nanomedicine in precision-targeted drug delivery systems [69, 70]. Furthermore, we have implemented an online calculator for acutely decompensated cirrhosis based on this stratification output, available at https://decision-for-liver.eu/for-scientists/clustall-web-application/.

Although our study showed promising results, it is important to acknowledge some limitations. Firstly, concerning our novel stratification framework, we designed a method aimed at minimizing user-defined parameters by exhaustively exploring all potential clustering solutions across various parameter combinations. However, practical decisions were made, such as employing PCA to diminish the dimensionality of highly correlated variables. In future iterations, we intend to explore alternatives such as Independent Component Analysis or PCA tailored for ordinal variables. Additionally, the determination of the number of components included in each dimensionality reduction will be guided by data-driven criteria. Furthermore, the ClustALL framework offers scope for expansion by incorporating additional methods and distance metrics, affording users the autonomy to select those most suitable for their needs. Secondly, our stratification relied solely on routinely available clinical data collected at hospital admission, potentially limiting the comprehensive understanding of patients' conditions. Future investigations should integrate biological data, preferably derived from multiomic analyses. It is also relevant to note that in the ACLARA cohort, predictive power was assessed only at the 28-day mark due to study design constraints. Moreover, it is worth highlighting that broader utilization of ClustALL (e.g., in other

complex diseases and/or including omic data) may shed light on areas necessitating refinement, aligning with the *No-Free Lunch* theorem discussed previously [21].

In summary, this study introduces a novel unsupervised clustering framework, ClustALL, capable of overcoming the limitations of available stratification methods. Expanding beyond cirrhosis, ClustALL shows potential for wider implementation across various clinical settings, hinting at its ability to transform patient subgroup identification, expand possibilities of drug repurposing [71], and in general, to enhance healthcare management.

Conclusions

ClustALL stands out as a comprehensive and versatile computational framework for unsupervised patient stratification that uses multimodal clinical data such as biochemical markers, demographics, and clinical scores as input. Furthermore, ClustALL ensures the identification of robust stratifications—including two concepts of robustness—and allows the identification of multiple robust stratifications over the same population. In the context of acute decompensation of cirrhosis, ClustALL not only successfully navigates the intricacies of diverse clinical information but also identifies several robust stratifications. Furthermore, validating findings across different time points and in an independent cohort underscores the reliability of ClustALL. Overall, this work not only contributes to our understanding of patient heterogeneity in cirrhosis but also positions ClustALL as a powerful stratification tool that could be applied to other diseases, thereby advancing precision medicine and facilitating the development of more targeted and effective clinical interventions. Future developments of the tool could expand ClustALL framework by incorporating biological data from multiomic analyses and offering further customizable user functions.

Abbreviations

ACLF	Acute-on-chronic liver failure
AD	Acute decompensation of cirrhosis
AIC	Akaike information criterion
AUC	Area under the curve
BIC	Bayesian Information Criterion (BIC)
BH	Benjamini and Hochberg
CRP	C-reactive protein
EOS	End of study
ER	Error rate
FN	False negatives
FP	False positives
INR	International Normalized Ratio
kNN	k-nearest neighbors
MICE	Multivariate imputation by chained equations
PC	Principal component
PCA	Principal component analysis

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-024-05386-2.

Supplementary Material 1

Supplementary Material 2

Supplementary Material 3

Supplementary Material 4

Supplementary Material 5

Supplementary Material 6

Acknowledgements

CANONIC investigators, PREDICT investigators, ACLARA investigators, and all DECISION investigators.

Author contributions

SPE, JesT, NK, VL, NP, and DGC conceived and designed the project. SPE, NK, NP, and DGC developed methodology. SPE, EH, AOL, FA, NP, and DGC analysed and interpreted data (statistical analysis, biostatistics, computational analysis). EMUR, FA, CPR, and CSG provided technical and material support (i.e., reporting or organizing data and constructing databases). SPE, EH, NP, PE, and DGC wrote the manuscript. JF, JC, PC, JT, and PE provided clinical interpretation of the results. EMUR, CPR, CLV, CA, WL, AQF, RM, JF VA, PC, CSG, JC, JT reviewed the manuscript.

Funding

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 847949. This research was made possible through access to data generated by the PREDICT and ACLARA studies promoted and funded by the European Foundation for the Study of Chronic Liver Failure (EF CLIF), a private, non-profit organization receiving unrestricted grants from Grifols and Fundació Privada Cellex. JT was appointed as visiting Professor in EF-Clif for the execution of the study by a grant from Cellex Foundation. JT was supported by the German Research Foundation (DFG) project ID 403224013 – SFB 1382 (A09), by the German Federal Ministry of Education and Research (BMBF) for the DEEP-HCC project and by the Hessian Ministry of Higher Education, Research and the Arts (HMWK) for the ENABLE and ACLF-I cluster projects. The MICROB-PREDICT (project ID 825694), DECISION (project ID 847949), GALAXY (project ID 668031), LIVERHOPE (project ID 731875), and IHMCSA (project ID 964590) projects have received funding from the European Union's Horizon 2020 research and innovation program. N-A.K and C.L-V were supported by DECISION (project ID 847949). N.P.P was funded by a Ramón y Cajal fellow (RYC2021-032197-I) from the MCIN/AEI/https://doi.org/10.13039/501100011033 and European Union "NextGenerationEU"/PRTR and by a Juan de la Cierva-formación fellow (FJC2019-042304-I) from the Spanish Ministry of Science and Innovation (MCIN). V.L was supported by funding from the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). J.C. laboratory is supported by MCIN/AEI, PID2022-138970OB-I00 10.13039/501100011033/FEDER, UE. P-E.R.'s research laboratory is supported by the Foundation pour la Recherche Médicale (FRM EQU202303016287), "Institut National de la Santé et de la Recherche Médicale" (ATIP AVENIR), the "Agence Nationale pour la Recherche" (ANR-18-CE14-0006-01, RHU QUID-NASH, ANR-18-IDEX-0001, ANR-22-CE14-0002) by « Émergence, Ville de Paris », by Fondation ARC and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 847949.

Data availability

Researchers who provide a methodology sound proposal can apply for the data, as far as the proposal is in line with the research consented by the patients. These proposals should be requested through https://www.clifresearch.com/decision/Home.aspx. Data requestors will need to sign a data transfer agreement. The code to generate the ClustALL method is available on GitHub, at https://github.com/TranslationalBioinformaticsUnit/ClustALL_AD/.

Declarations

Competing Interests

JT has received speaking and/or consulting fees from Versantis, Gore, Boehringer-Ingelheim, Falk, Grifols, Genfit and CSL Behring. WL received funding from the MICROB-PREDICT (project ID 825694) and consultant for the LIVERHOPE (project ID 731875).

Author details

[1]Unit of Translational Bioinformatics, Navarrabiomed - Fundación Miguel Servet, Pamplona, Spain
[2]European Foundation for the Study of Chronic Liver Failure, Barcelona, Spain
[3]Biochemistry and Molecular Genetics Service, Hospital Clínic-IDIBAPS, Barcelona, Spain
[4]CIBERehd, Barcelona, Spain
[5]Department of Biomedical Sciences, University of Barcelona, Barcelona, Spain
[6]Division of Gastroenterology and Hepatology, A.O.U. Città della Salute e della Scienza di Torino, Torino, Italy
[7]Department of Gastroenterology & Hepatology, Section of Liver & Biliopancreatic disorders and Liver Transplantation, University Hospitals Leuven, KU LEUVEN, Leuven, Belgium
[8]Department of Gastroenterology, Hospital das Clínicas, University of São Paulo School of Medicine, Paulo School, Brazil
[9]Université Paris-Cité, Inserm, Centre de recherche sur l'inflammation, UMR 1149, Paris, France
[10]Assistance Publique-Hôpitaux de Paris (AP-HP), Paris, France
[11]Hôpital Beaujon, Service d'Hépatologie, Clichy, France
[12]Department of Medical and Surgical Science, University of Bologna, Bologna, Italy
[13]Biological and Environmental Science and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[14]SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence, Thuwal, Saudi Arabia
[15]Institute of Chemical Biology, Ilia State University, Tbilisi 0162, Georgia
[16]Department of internal medicine B, University of Münster, Münster, Germany
[17]Algorithmic Dynamics Lab, Center for Molecular Medicine, Karolinska Institutet, Solna, Sweden
[18]Computational Biology Program, Universidad de Navarra, CIMA, Instituto de Investigación Sanitaria de Navarra (IdiSNA), Navarra 31008, Spain
[19]AP-HP, Hôpital Beaujon, Service d'Hépatologie, DMU DIGEST, Centre de Référence des Maladies Vasculaires du Foie, FILFOIE, ERN RARE-LIVER, Clichy, France
[20]Department of Oncology-Pathology, Karolinska Institutet, Solna, Sweden
[21]IRCCS Azienda Ospedaliera-Universitaria di Bologna, Bologna, Italy
[22]Unit of Computational Medicine, Department of Medicine, Center for Molecular Medicine, Karolinska Institutet, Karolinska University Hospital, Stockholm, Sweden
[23]Computer, Electrical and Mathematical Sciences and Engineering Division, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Received: 1 February 2024 / Accepted: 10 June 2024
Published online: 27 June 2024

References

1. Almendro V, Kim HJ, Cheng YK, Gonen M, Itzkovitz S, Argani P, et al. Genetic and phenotypic diversity in breast tumor metastases. Cancer Res. 2014;74(5):1338–48.
2. Kotelnikova E, Kiani NA, Abad E, Martinez-Lapiscina EH, Andorra M, Zubizarreta I et al. Dynamics and heterogeneity of brain damage in multiple sclerosis. PLoS Comput Biol. 2017;13(10).
3. Dennis JM, Shields BM, Henley WE, Jones AG, Hattersley AT. Disease progression and treatment response in data-driven subgroups of type 2 diabetes compared with models based on simple clinical features: an analysis using clinical trial data. Lancet Diabetes Endocrinol. 2019;7(6):442–51.
4. Schuppan D, Afdhal NH. Liver cirrhosis. Lancet. 2008;371(9615):838–51.
5. Mansour D, McPherson S. Management of decompensated cirrhosis. Clin Med (Lond). 2018;18(Suppl 2):s60–5.
6. D'Amico G, Morabito A, D'Amico M, Pasta L, Malizia G, Rebora P, et al. Clinical states of cirrhosis and competing risks. Journal of Hepatology. Volume 68. Elsevier B.V.; 2018. pp. 563–76.
7. Spach D. Evaluation and Prognosis of Patients with Cirrhosis - Core Concepts [Internet]. https://www.hepatitisC.uw.edu/go/evaluation-staging-monitoring/evaluation-prognosis-cirrhosis/core.
8. Cerezo Cerezo J, ALC. Population stratification: a fundamental instrument used for population health management in Spain: good practice brief. World Health Organization Regional Office for Europe; 2018. https://apps.who.int/iris/handle/10665/345586.
9. Moral TT, Sanchez-Niubo A, Monistrol-Mula A, Gerardi C, Banzi R, Garcia P, et al. Methods for stratification and validation cohorts: a scoping review. Volume 12. Journal of Personalized Medicine. MDPI; 2022.
10. Horne E, Tibble H, Sheikh A, Tsanas A. Challenges of clustering multimodal clinical data: review of applications in asthma subtyping. JMIR Medical Informatics. Volume 8. JMIR Publications Inc.; 2020.
11. Wang H, Donoho D, Kuppler C, Loftus TJ Jr, Copyright UG. frai, Phenotype clustering in health care: A narrative review for clinicians.
12. Saxena A, Prasad M, Gupta A, Bharill N, Patel OP, Tiwari A, et al. A review of clustering techniques and developments. Neurocomputing. 2017;267:664–81.
13. Cismondi F, Fialho AS, Vieira SM, Reti SR, Sousa JMC, Finkelstein SN. Missing data in medical databases: Impute, delete or classify? Artif Intell Med. 2013;58(1):63–72.
14. Rodríguez AH, Ruiz-Botella M, Martín-Loeches I, Jimenez Herrera M, Solé-Violan J, Gómez J et al. Deploying unsupervised clustering analysis to derive clinical phenotypes and risk factors associated with mortality risk in 2022 critically ill patients with COVID-19 in Spain. Crit Care. 2021;25(1).
15. Curtis JR, Weinblatt M, Saag K, Bykerk VP, Furst DE, Fiore S, et al. Data-Driven patient clustering and Differential Clinical outcomes in the Brigham and women's Rheumatoid Arthritis Sequential Study Registry. Arthritis Care Res (Hoboken). 2021;73(4):471–80.
16. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based Disease Risk Prediction. Front Bioinf. 2022;2.
17. Hennig C. What are the true clusters? Pattern Recognit Lett. 2015;64:53–62.
18. Lopez-Martinez-Carrasco A, Juarez JM, Campos M, Canovas-Segura B. A methodology based on Trace-based clustering for patient phenotyping. Knowl Based Syst. 2021;232.
19. Chalancon G, Kruse K, Babu MM. Clustering coefficient. Encyclopedia of systems Biology. New York, NY: Springer New York; 2013. pp. 422–4.
20. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
21. Adam SP, Alexandropoulos SAN, Pardalos PM, Vrahatis MN. In. No free lunch theorem: a review. 2019. p. 57–82.
22. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. Psychometrika. 1985;50(2):159–79.
23. Steinley D, Brusco MJ. Choosing the number of clusters in K-means clustering. Psychol Methods. 2011;16(3):285–97.
24. Altman N, Krzywinski M, Clustering. Nat Methods. 2017;14(6):545–6.
25. Kitano H. Towards a theory of biological robustness. Mol Syst Biol. 2007;3(1).
26. Yu H, Chapman B, Di Florio A, Eischen E, Gotz D, Jacob M, et al. Bootstrapping estimates of stability for clusters, observations and model selection. Comput Stat. 2019;34(1):349–72.
27. Lu Y, Phillips CA, Langston MA. A robustness metric for biological data clustering algorithms. BMC Bioinformatics. 2019;20(S15):503.
28. Trebicka J, Fernandez J, Papp M, Caraceni P, Laleman W, Gambino C, et al. The PREDICT study uncovers three clinical courses of acutely decompensated cirrhosis that have distinct pathophysiology. J Hepatol. 2020;73(4):842–54.
29. Schonlau M. Visualizing non-hierarchical and hierarchical cluster analyses with clustergrams. Vol. 19, Computational Statistics. 2004.
30. Ringnér M. What is principal component analysis? [Internet]. Vol. 26, NATURE BIOTECHNOLOGY. 2008. http://www.nature.com/naturebiotechnology.
31. Martínez-Gómez E, Richards MT, Richards DSP. DISTANCE CORRELATION METHODS FOR DISCOVERING ASSOCIATIONS IN LARGE ASTROPHYSICAL DATABASES. Astrophys J. 2014;781(1):39.

32. Gower JC. A General Coefficient of Similarity and Some of Its Properties. Vol. 27, Biometrics. 1971.
33. Hummel M, Edelmann D, Kopp-Schneider A. Clustering of samples and variables with mixed-type data. PLoS ONE. 2017;12(11).
34. Zhang Z, Murtagh F, Poucke S, Van, Lin S, Lan P. Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with R. Ann Transl Med. 2017;5(4).
35. Arora P, Deepali, Varshney S. Analysis of K-Means and K-Medoids Algorithm for Big Data. Physics Procedia. Elsevier B.V.; 2016. pp. 507–12.
36. Liu Y, Li Z, Xiong H, Gao X, Wu J. Understanding of internal clustering validation measures. In: Proceedings - IEEE International Conference on Data Mining, ICDM. 2010. pp. 911–6.
37. Zhao Q, Fränti P. WB-index: a sum-of-squares based index for cluster validity. Data Knowl Eng. 2014;92:77–89.
38. Fletcher S, Islam Z. Comparing sets of patterns with the Jaccard index. Volume 22. Australasian Journal of Information Systems Fletcher & Islam; 2018.
39. Tang M, Kaymaz Y, Logeman BL, Eichhorn S, Liang ZS, Dulac C, et al. Evaluating single-cell cluster stability using the Jaccard similarity index. Bioinformatics. 2021;37(15):2212–4.
40. Van Buuren S, Groothuis-Oudshoorn K. Journal of Statistical Software mice: Multivariate Imputation by Chained Equations in R [Internet]. Vol. 45. 2011. http://www.jstatsoft.org/.
41. Farias AQ, Curto Vilalta A, Momoyo Zitelli P, Pereira G, Goncalves LL, Torre A, et al. Genetic ancestry, race, and severity of acutely decompensated cirrhosis in Latin America. Gastroenterology. 2023;165(3):696–716.
42. Hennig C. Cluster-wise assessment of cluster stability. Comput Stat Data Anal. 2007;52(1):258–71.
43. R Core Team. R: a language and environment for statistical. Vienna, Austria: R Foundation for Statistical Computing; 2021.
44. Lagani V, Athineou G, Farcomeni A, Tsagris M, Tsamardinos I. Feature selection with the *R* Package MXM: discovering statistically equivalent feature subsets. J Stat Softw. 2017;80(7).
45. Tsagris M, Tsamardinos I. Feature selection with the R package MXM. F1000Res. 2018;7:1505.
46. Deng Z, Zhu X, Cheng D, Zong M, Zhang S. Efficient kNN classification algorithm for big data. Neurocomputing. 2016;195:143–8.
47. Ali N, Neagu D, Trundle P. Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. SN Appl Sci. 2019;1(12):1559.
48. Rossi R, Murari A, Gaudio P, Gelfusa M. Upgrading model selection criteria with goodness of fit tests for practical applications. Entropy. 2020;22(4):447.
49. Cook NR. Quantifying the added value of new biomarkers: how and how not. Diagn Progn Res. 2018;2(1):14.
50. Arroyo V, Moreau R, Jalan R. Acute-on-chronic liver failure. N Engl J Med. 2020;382(22):2137–45.
51. Stewart CA, Malinchoc M, Kim WR, Kamath PS. Hepatic encephalopathy as a predictor of survival in patients with end-stage liver disease. Liver Transpl. 2007;13(10):1366–71.
52. Jepsen P, Vilstrup H, Andersen PK. The clinical course of cirrhosis: the importance of multistate models and competing risks analysis. Hepatology. 2015;62(1):292–302.
53. D'Amico G, Morabito A, D'Amico M, Pasta L, Malizia G, Rebora P, et al. Clinical states of cirrhosis and competing risks. J Hepatol. 2018;68(3):563–76.
54. Castela Forte J, van der Yeshmagambetova G, Hiemstra B, Kaufmann T, Eck RJ et al. Identifying and characterizing high-risk clusters in a heterogeneous ICU population with deep embedded clustering. Sci Rep. 2021;11(1).
55. Li X, Wang C, Liu L, Xia X. A Method for Heterogeneity Analysis of Complex Diseases Based on Clustering Algorithm. In: Proceedings – 13th International Conference on Computational Intelligence and Security, CIS 2017. Institute of Electrical and Electronics Engineers Inc.; 2018. pp. 155–8.
56. Choobdar S, Ahsen ME, Crawford J, Tomasoni M, Fang T, Lamparter D, et al. Assessment of network module identification across complex diseases. Nat Methods. 2019;16(9):843–52.
57. Naithani N, Sinha S, Misra P, Vasudevan B, Sahu R. Precision medicine: Concept and tools. Med J Armed Forces India. 2021;77(3):249–57.
58. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nature Reviews Genetics. Volume 20. Nature Publishing Group; 2019. pp. 273–82.
59. Qi R, Ma A, Ma Q, Zou Q. Clustering and classification methods for single-cell RNA-sequencing data. Briefings in Bioinformatics. Volume 21. Oxford University Press; 2019. pp. 1196–208.
60. Coombes CE, Liu X, Abrams ZB, Coombes KR, Brock G. Simulation-derived best practices for clustering clinical data. J Biomed Inf. 2021;118.
61. Lu Y, Phillips CA, Langston MA. A robustness metric for biological data clustering algorithms. BMC Bioinformatics. 2019;20.
62. Müller E, Günnemann S, Färber I, Seidl T. Discovering multiple clustering solutions: grouping objects in different views of the data. In: Proceedings - International Conference on Data Engineering. 2012. pp. 1207–10.
63. Hu J, Pei J. Subspace multi-clustering: a review. Knowledge and Information Systems. Volume 56. Springer London; 2018. pp. 257–84.
64. Elkrief L, Rautou PE, Sarin S, Valla D, Paradis V, Moreau R. Diabetes mellitus in patients with cirrhosis: clinical implications and management. Liver Int. 2016;36(7):936–48.
65. Paternostro R, Jachs M, Hartl L, Simbrunner B, Scheiner B, Bauer D et al. Diabetes impairs the haemodynamic response to non-selective betablockers in compensated cirrhosis and predisposes to hepatic decompensation. Aliment Pharmacol Ther. 2023.
66. Romero-Gómez M, Montagnese S, Jalan R. Hepatic encephalopathy in patients with acute decompensation of cirrhosis and acute-on-chronic liver failure. J Hepatol. 2015;62(2):437–47.
67. Ferenci P. Hepatic encephalopathy. Gastroenterol Rep (Oxf). 2017;5(2):138–47.
68. Higuera-de-la-Tijera F, Velarde-Ruiz Velasco JA, Raña-Garibay RH, Castro-Narro GE, Abdo-Francis JM, Moreno-Alcántar R, et al. Current vision on diagnosis and comprehensive care in hepatic encephalopathy. Revista De Gastroenterología De México. (English Edition). 2023;88(2):155–74.
69. Khalilov RK. Future prospects of biomaterials in nanomedicine. Adv Biology Earth Sci. 2024;9(Special Issue):5–10.
70. Huseynov E. Novel nanomaterials for hepatobiliary diseases treatment and future perspectives. Adv Biology Earth Sci. 2024;9(Special Issue):81–91.
71. Ahmed F, Samantasinghar A, Soomro AM, Kim S, Choi KH. A systematic review of computational approaches to understand cancer biology for informed drug repurposing. J Biomed Inf. 2023;142:104373.

Publisher's Note