Analyzing trajectories of clinical markers in patients with sepsis through multivariate longitudinal clustering

Patrizia Ribino[1], Maria Mannone[1,2], Claudia Di Napoli[3], Giovanni Paragliola[3], Davide Chicco[4,5,6,*] and Francesca Gasparini[4,5]

[1]*Istituto di Calcolo e Reti ad Alte prestazioni, Consiglio Nazionale delle Ricerche (CNR), Palermo, Italy*
[2]*Institute of Physics and Astronomy, Universität Potsdam, Germany*
[3]*Istituto di Calcolo e Reti ad Alte prestazioni, Consiglio Nazionale delle Ricerche (CNR), Naples, Italy*
[4]*Dipartimento di Informatica Sistemistica e Comunicazione, Università di Milano-Bicocca, Milan, Italy*
[5]*NeuroMI, Milan Center for Neuroscience, Milan, Italy*
[6]*Institute of Health Policy Management and Evaluation, University of Toronto, Toronto, Ontario, Canada*

Abstract

Sepsis is a life-threatening condition with complex and dynamic progression, often requiring timely and personalized treatment strategies. In this paper, we propose a multivariate longitudinal clustering, an advanced data analysis technique, as a powerful approach to understanding the diverse trajectories of sepsis by grouping patients based on multiple clinical variables measured over time. Dynamic Time Warping (DTW) is integrated into the longitudinal clustering as a distance measure to identify subgroups of patients with similar temporal patterns in multivariate data. By leveraging sepsis-related electronic health records (EHRs), which provide rich time-series data on laboratory results along with patient demographics and underlying health conditions, the proposed method reveals distinct sepsis phenotypes that reflect variations in disease progression. Our results confirm the critical role of the Thrombin-Antigen complex and the International Normalized Ratio as predictors of poor outcomes for septic patients. Despite challenges like missing data and interpretability, multivariate longitudinal clustering in sepsis offers significant potential to enhance clinical decision-making and improve patient outcomes.

Keywords

clustering, unsupervised machine learning, electronic health records, longitudinal clustering, patient trajectories, sepsis, intensive care unit

1. Introduction

Sepsis is a complex, life-threatening condition caused by the body's overwhelming response to infection, often leading to multi-organ failure or septic shock [1]. Sepsis can arise from several types of infection, including bacterial, viral, fungal, or parasitic infections. Familiar sources of infection that lead to sepsis include Pneumonia, Urinary Tract Infections (UTIs), Abdominal infections, Bloodstream infections (bacteremia), and surgical site infections. In sepsis, the body's immune system triggers an excessive and harmful inflammatory response to fight infection. This can result in (i) Systemic Inflammatory Response Syndrome (SIRS) [2], namely a widespread release of inflammatory molecules that cause damage to blood vessels, leading to fluid leakage, edema, and reduced blood flow to vital organs, and (ii) Coagulopathy [3] caused by the activation of the clotting system, leading to the formation of small blood clots (microthrombi) in the microcirculation, which can further reduce blood flow and contribute to organ dysfunction. This can eventually lead to disseminated intravascular coagulation (DIC), characterized by excessive clotting and bleeding. Finally, Multiple organ dysfunction syndrome [4], due to the reduced blood flow and tissue damage caused by inflammation and micro-clot formation, can compromise the function of major organs, including heart, lungs, liver, kidneys, and brain.

HC@AIxIA 2024 – the 3rd AIxIA Workshop on Artificial Intelligence For Healthcare
*Corresponding author. Article version: 12th November, 2024 h15:47 CET.
davidechicco@davidechicco.it (Davide Chicco)
0000-0003-3266-9617 (Patrizia Ribino); 0000-0003-3606-3436 (Maria Mannone); 0000-0002-8626-5805 (Claudia Di Napoli); 0000-0003-3580-9232 (Giovanni Paragliola); 0000-0001-9655-7142 (Davide Chicco); 0000-0002-6279-6660 (Francesca Gasparini)

Electronic Health Records (EHRs) play a pivotal role in identifying, managing, and studying sepsis in clinical practice. Sepsis prediction models integrated into EHRs can exploit real-time patient data, and advanced machine learning (ML) techniques could enable early identification of patients at risk of developing sepsis, often before clinical signs become evident. These models have the potential to significantly improve sepsis outcomes by allowing for earlier interventions, such as antibiotic administration and fluid resuscitation, reducing the risk of severe complications or death. In [5], the authors developed a prediction model to evaluate the probability transition between three different disease states, estimating the daily evolution of disease severity during sepsis. XGBoost and LightGBM have been applied in [6] to predict early sepsis six hours in advance, while the authors in [7] used an LSTM network for the early detection of septic shock, showing that the proposed method allows one to detect patients up to 20 hours earlier.

Among the various ML techniques, unsupervised longitudinal clustering holds considerable potential for application within the context of electronic health records. The development of an EHR's longitudinal clustering allows for the grouping of patients according to their health data collected over time. Indeed, EHRs are longitudinal by nature: they capture patients' medical history, diagnoses, treatments, and outcomes over extended periods. Applying longitudinal clustering to these records involves analyzing patterns or trajectories within the data to identify subgroups of patients who exhibit similar health trends or progression of conditions [8].

Besides several works that apply static k-means to investigate phenotypic subgroups, including [9, 10, 11], several traditional unsupervised clustering models have been applied to group patients affected by sepsis, analyzing their data at specific time points. For instance, spectral clustering has been applied in [12] to find four groups of patients and study the trajectories of physiological variables associated with their risk-score clusters. Hierarchical clustering in [13] has been applied to identify clusters of patients according to clinical and biological characteristics collected at patients' admission. On the other hand, to our knowledge, only a few works adopted unsupervised longitudinal machine learning approaches. Among them, the authors in [14] and in [15] clustered groups of patients in heterogeneous medical conditions, using longitudinal k-means [16], and tracking longitudinal biomarkers, to understand the progression of Acute Kidney Injury and sepsis and their impact on mortality in patients with burns.

In this paper, we propose a multivariate longitudinal clustering for stratifying patients based on the similarity of their time-series profiles over the course of sepsis. We adopt Dynamic Time Warping (DTW) as a method for measuring the similarity between two time-series data points even if their trajectories differ in timing. The goal is to group patients based on the similarity of their sepsis-related trajectories rather than static measurements. For sepsis, DTW can align different patients' clinical variables, even if they deteriorate or recover at different rates. Patients in the same cluster are expected to have similar sepsis progression patterns across multiple variables. In contrast, those in different clusters might represent distinct clinical phenotypes of sepsis (e.g., fast vs. slow progression).

The rest of the paper is organized as follows. Section 2 introduces the materials and methods used in this work. Results and discussions are presented in Section 3. Finally, conclusions are drawn in Section 4.

2. Materials and Methods

2.1. Dataset

The dataset we analyze in this study was collected at the Jichi Medical University Hospital in Shimotsuke, Tochigi, Japan and was derived from data of patients admitted to the intensive care unit (ICU) of the university hospital between April 2014 and September 2016 [17]. The dataset was described by the data curators and released publicly in a 2018 study [17], following the FAIR (findability, accessibility, interoperability, and reusability) principles [18].

The original dataset contains data from 205 patients and includes both static (single-visit or status) features and longitudinal (multi-visit) variables. The static features indicate the sex of the patient, the source of sepsis (one among pulmonary, abdominal, urinary tract, soft tissue, bloodstream, or other),

and one or more comorbidities that could be present in addition to sepsis (ischemic heart disease, congestive heart failure, arrhythmia, chronic obstructive pulmonary disease, chronic kidney disease, and/or cardiovascular diseases). The dataset includes other static variables, such as APACHE II (Acute Physiology and Chronic Health Evaluation) score, which is commonly used in the ICU to predict outcomes in critically ill patients, including those with sepsis. The dataset considers a variety of physiological parameters (e.g., temperature, blood pressure, heart rate, oxygenation) along with chronic health conditions to estimate the risk of mortality.

The SOFA (Sequential Organ Failure Assessment) score is also present and includes INR (International Normalized Ratio) as a component to assess liver function, and coagulopathy is used to predict mortality in septic patients. Higher INR values contribute to a higher SOFA score, indicating a greater likelihood of death. Moreover, the dataset also incorporates the days spent by the patient in the ICU, if a patient died during the 28 days after the septic episode, and the number of days from the septic episode to the death (in case of death).

For the objective of this study, we do not consider both clinical variables strictly correlated to the prognosis and severity of sepsis to avoid bias in the results and static variables that are not useful in a longitudinal study. We only consider clinical variables that have a temporal dynamic and are monitored during the ICU stay. Hence, for each patient, the following sets of predictors from the first seven days after the start time of clinical concern were considered:

- The Thrombin-Antithrombin (TAT) complex is a biomarker that indicates the activation of the coagulation system. The TAT complex is often measured in clinical settings to assess the level of thrombin generation and overall coagulation activity;
- The International Normalized Ratio (INR) is a laboratory measurement used to assess the time it takes for blood to clot. It is a standardized Prothrombin Time (PT) version, allowing consistent results across different labs [19];
- The Fibrin Degradation Products (FDP) are fragments resulting from fibrin breakdown in blood clots. They are a crucial marker of fibrinolysis, which dissolves blood clots;
- The Absolute Immature Platelet Count (AIPC) measures the number of immature platelets in the bloodstream. Immature platelets are newly produced by the bone marrow and released into circulation. They reflect the bone marrow's platelet production rate and are often used as a marker of thrombopoiesis (platelet formation);
- Protein C (also called Anticoagulant Protein C, APC) is a vital anticoagulant protein in the body, playing a key role in regulating blood clotting and preventing excessive clot formation [20];
- Platelet Count (PlC) is a crucial component of a Complete Blood Count (CBC) that measures the number of platelets in a given blood volume. Platelets are small cell fragments that play a key role in blood clotting and wound repair by aggregating at the site of vascular injury and forming a clot.

2.2. Multivariate Longitudinal Cluster Analysis

The values of clinical variables over time for each patient can be seen as trajectories in a multidimensional space. We intend to cluster patients' trajectories according to their similarity and use this information to make predictions of medical interest. To this aim, we develop a k-means longitudinal clustering with soft Dynamic Time Warping (DTW) to discover patterns of joint trajectory in multivariate time series. Let us summarize the key ideas of the proposed method.

The first element of the algorithm is the well-known k-means. In a nutshell, a given dataset is partitioned into k parts, whose centroids are computed. Then, each data point is assigned to its closest cluster center, and the average of data points close to each centroid is considered a new set of centroids. The procedure is iterated until the algorithm converges toward an optimal cluster-center assignment [21, 22]. The second element of the algorithm is DTW, that is, an instance of shape-respecting distance. While the Euclidean distance fails to detect a similarity of shape between two curves, this information is caught by DTW. Considering as (discrete) curves the polylines joining the points of two time series,

respectively, the shift between them can occur over time. Considering time as the dependent variable, DTW finds the optimal correspondence between the two considered time series [23]. Here, we adopt *soft*-DTW, with a smoothed and differentiable cost function. In soft-DTW, all possible alignments, rather than the optimal one (as for DTW), are considered [24]. The smoothing degree is user-adjusted via the parameter γ. Such a technique shows a better performance for clustering. Furthermore, thanks to the differentiability of the soft-DTW cost function, this method can also be exploited for machine learning.

2.2.1. Cluster evaluation

The following clustering metrics have been considered to evaluate the performance of the longitudinal clustering. The Silhouette coefficient (S) [25], which measures the quality of cluster placement for each individual i, is computed as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \tag{1}$$

where $a(i)$ is the *tightness*, that is, the average distance from subject i to all other subjects in the same cluster, and $b(i)$ is the *degree of separation*, that is, the shortest average distance from subject i to any other cluster. The overall clustering quality is obtained by calculating the average silhouette score across all data points. The range of a silhouette coefficient is between -1 and 1.

The Calinski-Harabasz index (CHI) [26] assesses how well-defined and distinct the clusters are, according to their *intra-cluster* and *inter-cluster variability*, and is computed as:

$$CHI = \frac{BCSS/(k-1)}{WCSS/(n-k)}, \tag{2}$$

where WCSS is the sum of squared distances from each point to its cluster mean,BCSS is the sum of squared distances between each cluster mean and the overall mean, weighted by the number of data points in each cluster, n is the total number of data points, and k is the number of clusters. High CHI values indicate that the clusters are well-separated and compact. Low CHI values indicate that the clusters are less distinct and may overlap. A lower value suggests poor clustering performance.

Finally, the Davies-Bouldin Index [27] is a metric based on the average similarity ratio of each cluster with its most similar (i.e., closest) cluster, computed as follows:

$$DBI = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} R_{ij}, \tag{3}$$

where R_{ij} is the ratio of the inter-cluster distance to the intra-cluster distance, and k is the number of clusters.

While in the classical approach S, CHI, and DBI are evaluated according to the distance between data points, here, since we are working with longitudinal data, these indexes are evaluated with respect to the distance between a longitudinal patient profile and a set of multiple profiles.

3. Clustering Results

The multivariate longitudinal clustering method has been applied to the cohort of individuals defined in Section 2.1, to identify clinical sub-types in septic patients with different outcomes during ICU stays. Several tests were run by varying the number of clusters and the number and combinations of features. Table 1 reports the value of the Silhouette score, Caliski-Harabasz Index, and Davies-Bouldin Index related to the best configurations obtained by varying k and the feature combinations. As we can see, the best result was obtained by stratifying into two clusters and considering two features, the Thrombin-Antithrombin complex and the International Normalized Ratio.

Figure 1 illustrates the trajectories of the Thrombin-Antithrombin complex and International Normalized Ratio over the seven-day ICU period, corresponding to the patient profiles categorized within Cluster 0 and Cluster 1, found via the multivariate longitudinal cluster described in Section

Number of clusters = 2				
# Features	**Features Name**	**Silhouette Score**	**Calinski-Harabasz Index**	**Davies-Bouldin Index**
2	TAT - INR	0.85***	199.98***	0.57***
3	TAT - INR - FDP	0.52	53.04	1.3
4	TAT - INR - FDP - PlC	0.67	172.83	0.8
5	TAT - INR - PlC - AIPC - ProteinC	0.32	30.66	1.8
6	TAT - INR - PlC - FDP - AIPC - ProteinC	0.26	27.99	1.87
Number of clusters = 3				
# Features	**Features Name**	**Silhouette Score**	**Calinski-Harabasz Index**	**Davies-Bouldin Index**
2	INR-FDP	0.82	231.02	0.81
3	TAT - INR - FDP	0.55	104.75	1.0
4	TAT - INR - FDP - ProteinC	0.47	114.10	1.11
5	TAT - INR - PlC - AIPC - FDP	0.39	39.27	1.37
6	TAT - INR - PlC - FDP - AIPC - ProteinC	0.29	90.61	1.29
Number of clusters = 4				
# Features	**Features Name**	**Silhouette Score**	**Calinski-Harabasz Index**	**Davies-Bouldin Index**
2	TAT - INR	0.65	137.05	1.01
3	TAT - INR - FDP	0.55	64.78	1.67
4	TAT - INR - FDP - ProteinC	0.43	60.72	1.45
5	INR - PlC - AIPC - FDP- ProteinC	0.31	15.65	2.72
6	TAT - INR - PlC - FDP - AIPC - ProteinC	0.27	17.29	2.09

Table 1
Clustering results for $k = 2, 3, 4$. Silhouette score interval: $[-1; +1]$, higher the better. Calinski-Harabasz index interval: $[0; +\infty)$, the higher the better. Davies-Bouldin index interval: $[0; +\infty)$, the lower the better. *** best result for each index.

2.2. The observed trend indicates that ICU patients categorized within Cluster 0 exhibit, on average, elevated levels of Thrombin-Antithrombin Complex and International Normalized Ratio upon admission compared to individuals in Cluster 1. The INR values remained elevated for several days during the patient's stay in the intensive care unit. Then, these values deteriorated to a state deemed irreversible.

Conversely, ICU patients in Cluster 1 exhibit a more stable trajectory characterized, on average, by lower values than individuals in Cluster 0.

The characteristics of the cluster, including its size, average age, clinical variables, and the two primary patient outcomes of most relevant significance, are presented in Table 2. In such a table, a parameter called *28-day death* is also reported. It refers to the occurrence of death within 28 days of a specific event or condition, that is, sepsis in our case study. The distinctions among the clusters are further elucidated through the application of statistical methods, which assess their levels of statistical significance. To examine continuous data in relation to the normality of distributions, either analysis of variance (ANOVA) or the Kruskal-Wallis test was utilized. The chi-square test was exploited to assess the differences in the frequencies of categorical data. The significance level was set at $pvalue < 0.01$.

As we can see, there is no statistical difference in age and gender between Cluster 0 and Cluster 1. There are statistically more individuals with sepsis caused by pulmonary infections than other sources of sepsis in Cluster 0. The significant differences between patients in Cluster 0 and patients in Cluster 1 are related to the severity of their conditions. Patients in Cluster 0 show the worst APACHE II score and SOFA score with respect to patients in Cluster 1 ($pvalue = 4.6E-05$ and $pvalue = 1.9E-03$). The most noteworthy finding was that no individuals classified within Cluster 0 survived beyond 28 days. The 28-day death measure indicates that patients in Cluster 1 have a better chance of survival than patients in Cluster 0 ($pvalue = 1.3E-18$). The duration between the patient's admission to the intensive care unit and the event's occurrence was remarkably short for patients in Cluster 0: five days on average. On

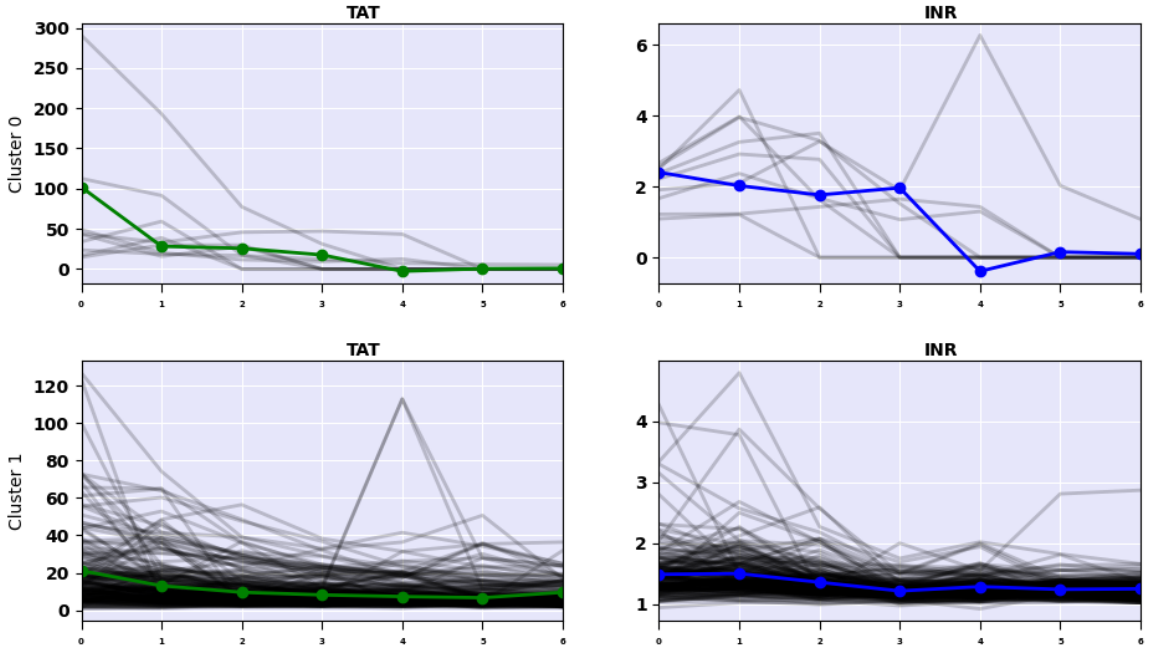

Figure 1: Joint progression of Thrombin-Antithrombin Complex and International Normalized Ratio for patients in Cluster 0 and Cluster 1

the contrary, the event for patients in Cluster 1 occurred on average after 20 days of ICU admission.

3.1. Discussions of the clustering results

This study focused on leveraging temporal clinical markers for stratifying septic patients according to their health status progression. By analyzing the longitudinal clinical variable measurements using an unsupervised clustering approach, which was opportunely developed for studying the temporal dynamics of patient health-status changes, we identified two groups of septic patients, where TAT and INR were identified as the two clinical variables most correlated to the different prognoses of ICU patients into different clusters.

Our results confirm that the Thrombin-Antithrombin Complex [20] and the International Normalized Ratio [19] are two biomarkers crucial in assessing and predicting sepsis progression. These markers help monitor coagulation and liver function, which are commonly affected in sepsis. Indeed, the Thrombin-Antithrombin Complex is a biomarker that indicates activation of coagulation. Thrombin is a key enzyme in blood clotting, while Antithrombin inhibits thrombin to prevent excessive clotting. When thrombin is activated during coagulation, it binds to antithrombin, forming the TAT complex. Elevated TAT levels in sepsis indicate hypercoagulation and may suggest the onset of Disseminated Intravascular Coagulation (DIC) [28], a serious complication of sepsis in which widespread clotting and bleeding occur simultaneously. As the medical literature reveals, high TAT levels are often correlated with poor outcomes in septic patients, including a higher risk of organ dysfunction and mortality. Moreover, TAT elevation is associated with increased thrombotic complications and coagulopathy, which are risk factors for multi-organ failure.

On the contrary, INR measures blood coagulation calculated from the prothrombin time (PT), which measures how long it takes to clot [19]. It is often used to assess the coagulation pathway and monitor patients on anticoagulant therapy. Sepsis affects coagulation by promoting excessive clot formation and reducing the blood's ability to clot when needed. This leads to an imbalance between procoagulant and anticoagulant factors. Elevated INR indicates impaired coagulation and suggests a deficiency in clotting factors. Higher INR values are associated with increased bleeding risk and are often a sign of liver dysfunction in septic patients. Elevated INR is a predictor of worse outcomes in sepsis, especially when it is accompanied by organ failure or the development of DIC. In sepsis, INR elevation may signal

Variables	Cluster 0 ($N = 9$)	Cluster 1 ($N = 195$)	p-value
Age (years)			3.1E-01†
mean ± SD	64.9 ± 12.9	68.5 ± 14.6	
[min, max]	[34, 81]	[19,101]	
Gender			1.6E-01‡
Male (%)	3 (33.3%)	114 (58.5%)	
Female (%)	6 (66.7%)	81 (41.5%)	
APACHE II score			4.6E-05†
mean ± SD	39.3 ± 8.8	24.6 ± 7.5	
[min, max]	[30, 53]	[7, 42]	
SOFA score			1.9E-03*
mean ± SD	11.7 ± 3.9	8 ± 3.3	
[min, max]	[4, 17]	[2, 19]	
ICU days			3.9E-04†
mean ± SD	4.6 ± 4.8	10.9 ± 7.7	
[min, max]	[2, 17]	[2, 58]	
28-day death			1.3E-18‡
survived (%)	0 (0%)	182 (93.3%)	
death (%)	9 (100%)	13 (6.7%)	
Time to event			8.6E-04†
mean ± SD	5.3 ± 6.5	19.9 ± 5.1	
[min, max]	[2, 22]	[11, 27]	
Source of sepsis			4.6E-60‡
Pulmonary (%)	6 (66.7%)	44 (22.6%)	
Abdominal (%)	2 (22.2%)	99 (50.8%)	
Urinary tract (%)	0 (0%)	11 (5.6%)	
Soft tissue (%)	0 (0%)	26 (13.3%)	
Blood stream (%)	0 (0%)	2 (1%)	
other (%)	1 (11.1%)	13 (6.7%)	

Table 2
Demographics and clinical characteristics of ICU patients at the admission. * One-way ANOVA, † Kruskal-Wallis test, ‡chi-square test

progression toward severe sepsis or septic shock, where coagulopathies and multi-organ failure are common.

Since TAT and INR are critical markers of sepsis coagulation dysfunction, their combined assessment can provide valuable insight into a patient's risk of progressing to more severe stages of sepsis, such as severe sepsis, septic shock, or multi-organ failure. As confirmed by our findings, a high TAT (indicating a pro-thrombotic state) and elevated INR (indicating impaired clotting) that has been present for days can indicate that a septic patient is at a critical point in disease progression and needs further urgent medical treatment.

4. Conclusions

The study proposed in this paper leverages the potentiality of longitudinal clustering to discover different profiles of septic patients, by including the temporal aspect of clinical variables delineating patients' health conditions. The main finding of this study is the distinction of patients with poorer prognoses within a brief timeframe of only a few days, by identifying the Thrombin–Antithrombin complex and International Normalized Ratio as the most critical clinical variables mainly correlated with adverse outcomes. This stratification has the potential to facilitate the development of tailored therapeutic strategies, such as implementing more aggressive care protocols designed explicitly for high-risk cohorts of septic patients.

Moreover, considering the high-dimensional characteristics of Electronic Health Records (EHR), particularly in the context of sepsis, which encompasses a diverse array of vital signs (such as heart rate, blood pressure, body temperature, and oxygen saturation), laboratory results (including white blood cell count, C-protein levels, platelet count, and Thrombin-Antithrombin complex, among others), as well as patient demographics (including age, sex, and underlying health conditions), the application of multivariate longitudinal clustering proves advantageous in discerning patient subgroups that may exhibit distinct patterns of sepsis progression. Hence, multivariate longitudinal clustering in the EHR of septic patients offers a powerful tool for better understanding patient health trajectories, supporting decision-making, and contributing to advances in personalized medicine.

However, the major limitation of this study is that the analytical framework relied exclusively on data derived from a single dataset. An analysis encompassing a larger dataset, and more datasets, may enhance the validity of the generalizations derived from our findings. In addition, the reported results have to be supported by clinicians to assess their clinical validity. For future work, we plan to develop different kinds of longitudinal clustering based on probabilistic approaches and fuzzy clustering. This could be more helpful when the boundaries between clusters are not well-defined, as may occur in medical settings.

List of abbreviations AIPC: absolute immature platelet counts. APACHE II: Acute Physiology and Chronic Health Evaluation II. CHI: Calinski-Harabasz index. CHF: congestive heart failure. CKD: chronic kidney disease. COPD: chronic obstructive pulmonary disease. CVD: cardiovascular disease. DBI: Davies-Bouldin index. DIC: disseminated intravascular coagulation. DTW: Dynamic Time Warping. EHR: electronic health record. FDP: fibrin degradation product. ICU: intensive care unit. IHD: ischemic heart disease. INR: international normalized ratio. LSTM: Long short-term memory. ML: machine learning. S: Silhouette. SIRS: Systemic Inflammatory Response Syndrome. SOFA: sequential organ failure assessment. TAT: Thrombin–Antithrombin complex. UTIs: Urinary tract infections; infections; DIC: Disseminated Intravascular Coagulation.

Conflict of interests The authors declare they have no conflict of interest.

Ethics approval and consent to participate The authorization for collecting the data from patients and releasing them publicly was obtained by the original dataset curators [17].

Code availability The code can be shared from the first author upon request.

Data availability The dataset is publicly available under the Attribution 4.0 International Deed (CC BY 4.0) license on Figshare at the following URL: https://figshare.com/articles/dataset/Time_course_of_immature_platelet_count_and_its_relation_to_thrombocytopenia_and_mortality_in_patients_with_sepsis/5837823?file=10343616

Funding This study work was funded by the European Union – Next Generation EU programme, in the context of The National Recovery and Resilience Plan, Investment Partenariato Esteso PE8 "Conseguenze e sfide dell'invecchiamento", Project Age-It (Ageing Well in an Ageing Society). This work was also partially supported by Ministero dell'Università e della Ricerca of Italy under the "Dipartimenti di Eccellenza 2023-2027" ReGAInS grant assigned to Dipartimento di Informatica Sistemistica e Comunicazione at Università di Milano-Bicocca. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

[1] R. S. Hotchkiss, L. L. Moldawer, S. M. Opal, K. Reinhart, I. R. Turnbull, J.-L. Vincent, Sepsis and septic shock, Nature Reviews Disease Primers 2 (2016) 1–21.

[2] M. Davies, P.-O. Hagen, Systemic inflammatory response syndrome, British Journal of Surgery 84 (1997) 920–935.

[3] A. G. Tsantes, S. Parastatidou, E. A. Tsantes, E. Bonova, K. A. Tsante, P. G. Mantzios, A. G. Vaiopoulos, S. Tsalas, A. Konstantinidi, D. Houhoula, N. Iacovidou, D. Piovani, G. K. Nikolopoulos, R. Sokou, Sepsis-induced coagulopathy: an update on pathophysiology, biomarkers, and current guidelines, Life 13 (2023) 350.
[4] G.-D. Sun, Y. Zhang, S.-S. Mo, M.-Y. Zhao, Multiple organ dysfunction syndrome caused by sepsis: risk factor analysis, International Journal of General Medicine (2021) 7159–7164.
[5] P. M. Klein Klouwenberg, C. Spitoni, T. van der Poll, M. J. Bonten, O. L. Cremer, Predicting the clinical trajectory in critically ill patients with sepsis: a cohort study, Critical Care 23 (2019) 1–9.
[6] X. Zhao, W. Shen, G. Wang, Early prediction of sepsis based on machine learning algorithm, Computational Intelligence and Neuroscience 2021 (2021) 6522633.
[7] J. Fagerström, M. Bång, D. Wilhelms, M. Chew, Lisep lstm: A machine learning algorithm for early detection of septic shock, Scientific Reports 9 (2019) 15132. doi:10.1038/s41598-019-51219-4.
[8] P. Ribino, C. Di Napoli, G. Paragliola, L. Serino, F. Gasparini, D. Chicco, Exploratory analysis of longitudinal data of patients with dementia through unsupervised techniques., in: Proceedings of AIxAS – the 4th Italian Workshop on Artificial Intelligence for an Ageing Society co-located with 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023), Rome, Italy, 9th November 2023, volume 3623 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 67–87. URL: https://ceur-ws.org/Vol-3623/AIxAS_2023_paper_8.pdf.
[9] R. Balk, A. M. Esper, G. S. Martin, R. R. Miller III, B. K. Lopansri, J. P. Burke, M. Levy, R. E. Rothman, F. R. D'Alessio, V. K. Sidhaye, et al., Rapid and robust identification of sepsis using septicyte rapid in a heterogenous patient population, medRxiv (2024) 2024–08.
[10] T. Zhang, S. Wang, D. Hua, X. Shi, H. Deng, S. Jin, X. Lv, Identification of zip8-induced ferroptosis as a major type of cell death in monocytes under sepsis conditions, Redox Biology 69 (2024) 102985.
[11] C. J. knobloch, A. Bourbia, AI-driven sepsis mortality analysis: identifying phenotypic clusters using unsupervised machine learning, Chest 166 (2024) a6416.
[12] R. Liu, J. L. Greenstein, J. C. Fackler, M. M. Bembea, R. L. Winslow, Spectral clustering of risk score trajectories stratifies sepsis patients by clinical outcome and interventions received, Elife 9 (2020) e58142.
[13] G. Papin, S. Bailly, C. Dupuis, S. Ruckly, M. Gainnier, L. Argaud, E. Azoulay, C. Adrie, B. Souweine, D. Goldgran-Toledano, et al., Clinical and biological clusters of sepsis patients using hierarchical clustering, PloS one 16 (2021) e0252793.
[14] M. Kim, D. Kym, J. Hur, J. Park, J. Yoon, Y. S. Cho, W. Chun, D. Yoon, Tracking longitudinal biomarkers in burn patients with sepsis and acute kidney injury: an unsupervised clustering approach, European Journal of Medical Research 28 (2023) 295.
[15] J. Yoon, D. Kym, J. Hur, Y.-S. Cho, W. Chun, D. Yoon, Longitudinal profile of routine biomarkers for mortality prediction using unsupervised clustering algorithm in severely burned patients: a retrospective cohort study with prospectively collected data, Annals of surgical treatment and research 104 (2023) 126–135.
[16] C. Genolini, B. Falissard, Kml: A package to cluster longitudinal data, Computer methods and programs in biomedicine 104 (2011) e112–e121.
[17] K. Koyama, S. Katayama, T. Muronoi, K. Tonai, Y. Goto, T. Koinuma, J. Shima, S. Nunomiya, Time course of immature platelet count and its relation to thrombocytopenia and mortality in patients with sepsis, PLOS One 13 (2018) e0192064. URL: https://doi.org/10.1371/journal.pone.0192064.
[18] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR guiding principles for scientific data management and stewardship, Scientific Data 3 (2016).

doi:10.1038/sdata.2016.18.
[19] A. Doargaleh, E. Favaloro, M. Bahraini, F. Rad, Standardization of Prothrombin Time/International Normalized Ratio (PT/INR), Int. J. Lab. Hematol. 43 (2021) 21–28.
[20] B. Dahlbäck, B. O. Villoutreix, The anticoagulant protein C pathway, FEBS Letters 15 (2005) 3310–3316. URL: https://doi.org/10.1378/chest.124.3_suppl.26S.
[21] K. Yang, M. Mohammadi Amiri, S. R. Kulkarni, Greedy centroid initialization for federated *k*-means. Knowledge and Information Systems, International Transactions in Operational Research (2024).
[22] C. Genolini, B. Falissard, Kml: A package to cluster longitudinal data, Computer Methods and Programs in Biomedicine 104 (2011) e112–e121. URL: https://doi.org/10.1016/j.cmpb.2011.05.008.
[23] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition, IEEE Transactions on Acoustics, Speech and Signal Processing 26 (1978) 43–49.
[24] M. Cuturi, M. Blondel, Soft-DTW: a differentiable loss function for time-series, in: Proceedings of the 34th International Conference on Machine Learning, PMLR 70, 2017, Sydney, Australia, 2017.
[25] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. doi:10.1016/0377-0427(87)90125-7.
[26] T. Calinski, J. Harabasz, A dendrite method for cluster analysis, Communications in Statistics – Theory and Methods 3 (1974) 1–27. doi:10.1080/03610927408827101.
[27] D. L. Davies, D. W. Bouldin, A cluster separation measure, IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI-1 (1979) 224–227. doi:10.1109/tpami.1979.4766909.
[28] K. Adelborg, J. L. Larsen., A.-M. Hvas, Disseminated intravascular coagulation: epidemiology, biomarkers, and management, British Journal of Haematology 192 (2021) 803–818. URL: https://doi.org/10.1111/bjh.17172.