Journal of Biomedical Informatics 152 (2024) 104615

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original Research

Soft phenotyping for sepsis via EHR time-aware soft clustering

Shiyi Jiang [a,1], Xin Gai [b,1], Miriam M. Treggiari [c,2], William W. Stead [d], Yuankang Zhao [e], C. David Page [e], Anru R. Zhang [e,f,*,2]

[a] Department of Electrical & Computer Engineering, Duke University, Durham, 27708, NC, USA
[b] Department of Statistical Science, Duke University, Durham, 27708, NC, USA
[c] Department of Anesthesiology, Duke University, Durham, 27708, NC, USA
[d] Department of Biomedical Informatics, Vanderbilt University, Nashville, 37235, TN, USA
[e] Department of Biostatistics & Bioinformatics, Duke University, Durham, 27708, NC, USA
[f] Department of Computer Science, Duke University, Durham, 27708, NC, USA

ARTICLE INFO

Dataset link: https://mimic.mit.edu, https://eicu-crd.mit.edu

Keywords:
Sepsis sub-phenotyping
EHR
Soft clustering
Semi-supervised learning

ABSTRACT

Objective: Sepsis is one of the most serious hospital conditions associated with high mortality. Sepsis is the result of a dysregulated immune response to infection that can lead to multiple organ dysfunction and death. Due to the wide variability in the causes of sepsis, clinical presentation, and the recovery trajectories, identifying sepsis sub-phenotypes is crucial to advance our understanding of sepsis characterization, to choose targeted treatments and optimal timing of interventions, and to improve prognostication. Prior studies have described different sub-phenotypes of sepsis using organ-specific characteristics. These studies applied clustering algorithms to electronic health records (EHRs) to identify disease sub-phenotypes. However, prior approaches did not capture temporal information and made uncertain assumptions about the relationships among the sub-phenotypes for clustering procedures.
Methods: We developed a time-aware soft clustering algorithm guided by clinical variables to identify sepsis sub-phenotypes using data available in the EHR.
Results: We identified six novel sepsis hybrid sub-phenotypes and evaluated them for medical plausibility. In addition, we built an early-warning sepsis prediction model using logistic regression.
Conclusion: Our results suggest that these novel sepsis hybrid sub-phenotypes are promising to provide more accurate information on sepsis-related organ dysfunction and sepsis recovery trajectories which can be important to inform management decisions and sepsis prognosis.

1. Introduction

Sepsis is a life-threatening organ dysfunction syndrome secondary to a dysregulated host response to infection, and the primary cause of death from infection, especially if not recognized and treated promptly [1]. A hallmark of sepsis is the heterogeneity of its presentation and its prognosis, due to the variability in pathogen and immune host response interactions.

In 2016, a consensus conference provided an updated definition of sepsis, with septic shock representing a subset of sepsis in which particularly profound circulatory, cellular, and metabolic abnormalities lead to substantially increased mortality [1].

The consensus definition emphasized the importance of timely recognition and prompt management of sepsis [2]. Available therapies and management for patients with sepsis remain limited to source control, administration of antibiotics, and supportive care [3]. Accumulated evidence suggests that the intrinsic heterogeneity of sepsis and variable stage at presentation posed challenges not only to clinical care but also to the conduct of clinical trials assessing interventions for sepsis. Therefore, identifying its sub-phenotypes is crucial for informing prognostic assessment and developing and evaluating effective treatment plans.

A prior study identified sepsis phenotypes at the time of patient presentation to the emergency department, using only routinely available Electronic Health Record (EHR) data in the clustering models [4]. The phenotypes were derived from a large observational cohort to ensure generalizability. This important study, however, did not account for the temporal registration and the rapidly evolving changes in patient physiological and laboratory values. Information acquired in the early

* Corresponding author at: Department of Biostatistics & Bioinformatics, Duke University, Durham, 27708, NC, USA.
E-mail address: anru.zhang@duke.edu (A.R. Zhang).
[1] These authors contributed equally to this work.
[2] These authors contributed equally to this work.

https://doi.org/10.1016/j.jbi.2024.104615
Received 1 December 2023; Received in revised form 25 January 2024; Accepted 20 February 2024
Available online 27 February 2024

course of sepsis can substantially enrich the clinical phenotypes, enable the identification of sub-phenotypes, and increase prognostic accuracy. Other studies have captured the dynamic nature of the clinical course in patients with sepsis using the change in the Sequential Organ Failure Assessment (SOFA) score that assesses the severity of organ dysfunction in ICU patients [5]. However, these scores have been used primarily as outcome measures to evaluate the overall course of organ dysfunction and to predict mortality.

To further advance the classification of sepsis, and identify potential subgroups, we incorporated medical context and temporal biomarker characteristics into the sepsis classification algorithms, early after sepsis onset.

Researchers have been studying disease phenotyping with the help of machine learning techniques and Electronic Health Records (EHRs) [6–10], which contain large amounts of patient-level information, including demographics, vital signals, lab tests, medications, and diagnosis. However, in recent review papers, Yang et al. and He et al. [11,12] pointed out that most existing literature used purely data-driven approaches and seldom considered real-world medical use cases and corresponding medical interpretations. Limited work considers temporal information in the EHR longitudinal data. In addition, few existing studies perform non-overlapping clustering, i.e., each patient is commonly assigned to only one group (sub-phenotype).

Sepsis may initially be associated with dysfunction of one organ system and progress to involve multiple organ systems. Because of the involvement of multiple systems, a patient may exhibit more than one sub-phenotype. We thus develop a soft clustering method that allows each patient to be assigned to more than one sub-phenotype. At the same time, we take biomarker temporal information into account and incorporate clinical information into the soft clustering algorithm. By applying transformations to the soft clustering results, we obtain six novel sepsis hybrid sub-phenotypes. We evaluate the plausibility of the results by providing a biological explanation. Additionally, built upon the soft clustering results, we train and validate a sepsis early-warning model to predict the novel sepsis hybrid sub-phenotypes. The results suggest the newly identified hybrid sub-phenotypes provide characterizations of different sepsis progressions.

Summary	Description
Problem	Due to the heterogeneity of sepsis, there are limitations on current sepsis characterization and subsequent patient treatment and management plans. Identifying novel sepsis sub-phenotypes is thus crucial for tackling these limitations.
What is Already Known	Limited literature on disease phenotyping using the EHR data considers temporal data. Additionally, they use a data-driven approach without accounting for medical context and make clinically arguable assumptions on the relationships between the sub-phenotypes.
What this Paper Adds	This study proposes a novel pipeline that combines computational models with temporal biomarker data and clinical context, which has not yet received much attention in the field of disease phenotyping. Additionally, this framework can be easily extended for other disease phenotyping.

2. Background and significance

2.1. Disease sub-phenotyping using EHR

With the growing resource of EHR availability, researchers began to identify disease sub-phenotypes using EHR to better characterize the diseases and provide insights for subsequent treatment plans. Wang et al. [13] proposed an algorithm that is built upon Latent Dirichlet Allocation for topic modeling to identify latent patient subgroups from three patient cohorts. Ibrahim et al. [14] utilized hierarchical clustering to identify sepsis sub-populations. Oh et al. [15] applied agglomerative hierarchical clustering to identify COVID-19 sub-phenotypes. Seymour et al. [4] discovered four novel sepsis clinical phenotypes by applying consensus K-Means clustering. However, none of the prior work utilizes the temporal information contained in the EHR. They typically used representative values at a certain time point within a defined time range, failing to capture changes in feature patterns through time.

There is a limited amount of work that considers temporal information. For instance, Xu et al. [16] transformed acute kidney injury (AKI) EHR longitudinal data into vector representations using memory networks and performed K-Means clustering after applying dimensional reduction to the transformed data. They identified three novel AKI sub-phenotypes with distinct characteristics. Lasko and Mesa [17] transformed longitudinal EHR data into continuous space and applied independent components analysis to identify sub-phenotypes of liver diseases. Smith et al. [18] proposed an algorithm to detect sepsis patients using longitudinal EHR data via Jensen–Shannon Divergence. Estiri et al. [19] transformed medication and diagnosis records into vectors and performed semi-supervised learning for phenotyping. Lee and Schaar [20] developed a dynamic clustering algorithm using deep learning for phenotyping. However, they utilized data-driven approaches and did not incorporate medical context into the designed models. We thus propose a soft clustering algorithm integrated with medical context to better characterize disease sub-phenotypes.

2.2. Soft clustering algorithms

Clustering is an important group of unsupervised learning algorithms that groups data samples based on similarity with a wide range of applications, such as biomedical data analysis, anomaly detection, and building recommendation systems [21]. Conventional clustering algorithms, such as K-Means clustering [22], hierarchical clustering [23] and DBSCAN [24], assign one sample to exclusively one cluster. Such algorithms are termed hard clustering.

Correspondingly, another category of algorithms that allows one sample to be assigned to multiple clusters is termed soft clustering algorithms. For instance, Fuzzy C-Means (FCM) clustering [25] is an algorithm that is built upon the K-Means clustering that assigns each sample with a degree of membership to each cluster. The degree of memberships to all clusters adds up to one. Cleuziou [26] proposed overlapping K-Means (OKM) clustering that assigns one sample to multiple clusters. Rather than expressing cluster assignment as degrees of memberships, OKM uses a set to indicate cluster assignment to each sample, where the sample is either a non-member or a member of the cluster. However, both FCM and OKM are sensitive to cluster centroid initialization and can be easily affected by outliers. Zhang et al. proposed K-Harmonic Means (KHM) clustering [27] that utilizes the harmonic average to address the algorithm instability due to different cluster centroid initialization. There exist many other soft clustering methods that are introduced in the survey from Ferraro and Giordani [28]. To the best of our knowledge, there is no work that utilizes soft clustering methods to tackle sepsis sub-phenotyping using EHR data.

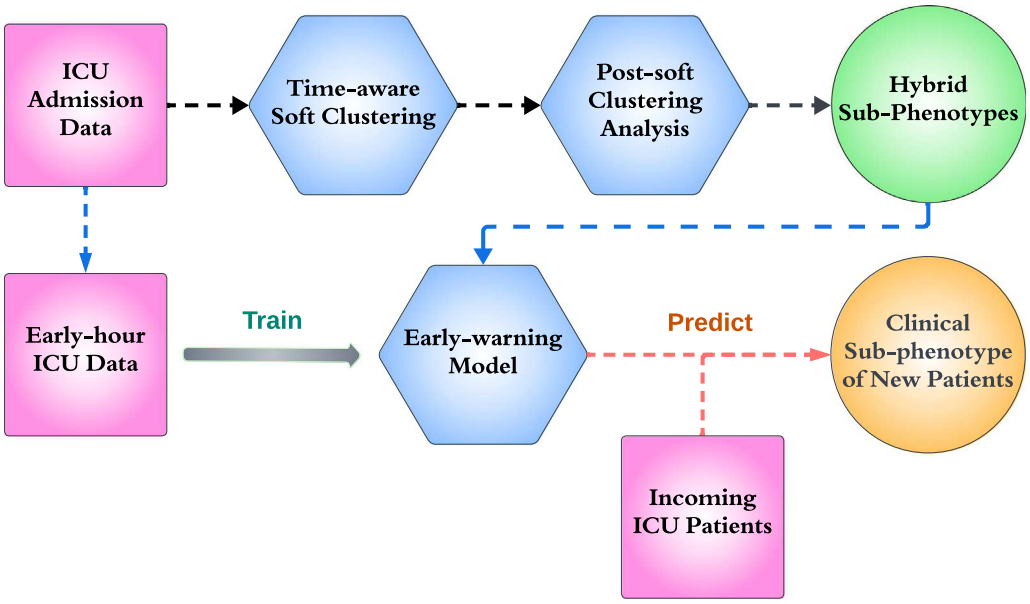

Fig. 1. Overall framework of the proposed method. Pink squares indicate data, blue hexagons represent algorithms/models, and circles in green and gold describe the outcome at the training and prediction stages, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Materials and methods

In this section, we provide a description of the datasets utilized in this study and the data selection/preprocessing steps. We next explain in detail each module of the proposed method for sepsis phenotyping illustrated in Fig. 1. We develop a time-aware soft-clustering algorithm followed by post-soft clustering analysis to identify potential novel sepsis sub-phenotypes. After careful evaluation, the resulting novel sub-phenotypes are utilized for sepsis early-warning prediction.

3.1. Data

3.1.1. Medical Information Mart for Intensive Care (MIMIC)-IV database

The MIMIC-IV database contains de-identified medical information on over 40,000 patients admitted to the intensive care units (ICU) of the Beth Israel Deaconess Medical Center (BIDMC) from 2008 to 2019. It contains information from many aspects, such as demographics, admissions, vital signs, laboratory tests, diagnosis, and treatments.

3.1.2. eICU collaborative research database

The eICU database contains medical records of over 200,000 patients admitted to the ICU in the continental US collected in 2014 and 2015. Similar to the MIMIC-IV database, the eICU includes information about patient demographics, admissions, diagnosis, medications, laboratory tests, etc.

3.2. Cohort selection and preprocessing

We extracted patient data based on Diagnosis Related Group (DRG) Codes [29], which are classified based on the International Classification of Diseases (ICD) diagnosis [30], age, sex, surgical procedures, discharge status, and comorbidity. We selected patients with DRG codes 870 (septicemia or severe sepsis with mechanical ventilation >96 h), 871 (septicemia or severe sepsis without mechanical ventilation >96 h with major complication or comorbidity), and 872 (septicemia or severe sepsis without mechanical ventilation >96 h without major complication or comorbidity). Since the eICU dataset does not contain DRG codes, we used the corresponding ICD codes that are mapped to the DRG codes according to [31]. Based on the characteristics of the sepsis [32], we chose records of the first 120 h of the last ICU stay from each patient. We chose variables included in or that contribute to the SOFA score because the third international consensus definitions for sepsis and septic shock (Sepsis-3) considers changes in SOFA as an indicator of sepsis progression [1]. We thus selected the following features: Arterial Blood Pressure systolic, Base Excess, Creatinine, Heart Rate, International Normalized Ratio of Prothrombin Time (PT-INR), Lactate, and Respiratory Rate.

We format the data into the form of a tensor $\mathcal{D} \in \mathbb{R}^{N\times P\times T}$, where N is the number of subjects, P is the number of clinical features, and T is the number of time steps in hours. Since the raw data contains a large amount of missing data, we conduct missing data imputation by introducing a novel method that uses a combination of low-rank matrix completion [33] and EHR timeline registration from our prior work [34]. We formulate data imputation as an alternating minimization problem. The goal is to find $U \in \mathbb{R}^{N\times r}$, $V \in \mathbb{R}^{PT\times r}$, and τ, such that the following objective function in Eq. (1) is minimized:

$$\min_{U,V,\tau} \sum_{i=1}^{N}\sum_{j=1}^{PT} A_{ij}\cdot\left(D_\tau[i,j]-U_{[i,:]}V_{[j,:]}^{\top}\right)^2. \tag{1}$$

Here, r is the matrix rank, τ is the discrete amount of time shift and $A \in \mathbb{R}^{N\times PT}$ is the indicator for the missing data. The alternating minimization contains two steps: (1) Obtain U and V from low-rank matrix completion on $D_\tau \in \mathbb{R}^{N\times PT}$ while τ is fixed. Note that D_τ is a matrix with D shifted by τ and reshaped into the dimension of $\mathbb{R}^{N\times PT}$. (2) Find optimal shift τ while U and V are fixed. We repeat the above steps until convergence. The final imputed data tensor $\mathcal{D}$ is approximated by $U\cdot V^T$, reshaped to a tensor of dimension $\mathbb{R}^{N\times P\times T}$.

3.3. Time-aware soft clustering

We developed a time-aware soft clustering algorithm for EHR data inspired by the work of Khanmohammadi et al. [35]. It is based on a hybrid of the harmonic K-Means clustering algorithm and the overlapping K-Means clustering algorithm. This hybrid algorithm is less sensitive to the initial cluster centroids selection and thus has improved algorithm stability.

We present the proposed soft clustering algorithm in Algorithm 1 (see Appendix). We denote all data records from the selected cohort as $D = \{x_i, i = 1,\dots,N\}$, where $x_i \in \mathbb{R}^{P\times T}$ represents data of each subject i of the N total subjects (equivalent to $\mathcal{D}[i,:,:]$). To represent

clinical context, we denote L as a collection of binary vectors $l_i \in \mathbb{R}^3$, where each element in l_i indicates the existence of an organ dysfunction type on subject i. For sub-phenotyping, we selected organ dysfunctions representing the lung, liver, and kidney based on their contribution to the SOFA score computation. We create L based on groups of ICD-9 codes for the three types of organ dysfunction: liver (570.*–573.*), kidney (580.*–589.*), and lung-related (510.*–519.*) diseases [30]. For instance, if subject i is diagnosed with ICD-9 codes 573.9 and 584.9 but none from 510.* to 519.*, the corresponding l_i is $(1, 1, 0)^T$.

According to Basu et al. [36], labeled data used to develop initial cluster centroids and cluster constraints effectively enhances the quality and stability of the clustering result. We thus initialized K cluster centroids $\{temp_k, k = 1, \ldots, K\}$ using the average of subjects with each of the single types of organ dysfunction (i.e., one cluster centroid for each of the liver, kidney, and lung-only dysfunction subject groups). We then perform cluster assignments to each subject according to Algorithm 2 (see Appendix) based on the overlapping K-Means clustering. We compute the distance between the subject i and each cluster centroid and assign the subject to its nearest cluster centroid. Subsequent cluster assignments depend on $\Phi(x_i)$, the average of the assigned cluster centroids to subject i, and $\Phi(x_i)'$, the average of the assigned centroids and the nearest candidate centroid to subject i. If the subject is closer to $\Phi(x_i)'$ than to $\Phi(x_i)$, the individual is then assigned to the nearest candidate cluster centroid. After obtaining cluster assignments $\{m_i^{(0)}, i = 1, \ldots, N\}$, where each $m_i^{(0)}$ is a set of cluster membership indicators, we update the cluster centroids by applying transformations shown in Step 2 of the Algorithm 1. We iteratively update the cluster centroids and cluster assignments until convergence.

To incorporate clinical context into the algorithm, we employ semi-supervised learning that calibrates cluster centroids after updating cluster assignments at each iteration based on the ICD group information $L = \{l_i, i = 1, \ldots, N\}$ of each subject shown in Algorithm 3 (see Appendix). We compute a weighted sum of distances $unsupLoss$ for all subjects which is built upon the objective function of the fuzzy C-Means algorithm [25]. Additionally, we compute $supLoss$ to enforce each subject with a single ICD group label to be closer to the targeted cluster centroid and further away from the non-targeted cluster centroids. Note that we assume each cluster centroid represents a designated organ dysfunction type. We use scalar hyperparameters β_1 and β_2 to adjust the strength of the constraint within $supLoss$. β_1 controls $tLoss$, the degree to which each subject's (with a single ICD group labeled) distance to the targeted cluster centroid $temp_{l \in l_i}$. β_2 adjusts $ntLoss$, the degree to which each subject's distance to the non-targeted cluster centroids $\{temp_{l \notin l_i}\}$. Finally, we calibrate cluster centroids by applying the stochastic gradient descent (SGD) to the $totLoss$.

After the iterative updates and calibrations of the cluster centroids reach convergence, we output a distance matrix $\{d_{ik}\} \in \mathbb{R}^{N \times K}$, where each element d_{ik} indicates the distance between the subject i to the cluster centroid k. Note that when computing the distance of each subject to a cluster centroid, we only consider the first 24 h of data in the ICU for features including systolic blood pressure, base excess, and respiratory rate since the effects of treatments may affect subsequent data patterns for different phenotypes. The resulting distance for each above-mentioned feature was multiplied by five to ensure the computed distance for each feature having the same scale. We used entire 120-hour data to compute distance for the rest of the features. We then compute cluster membership matrix $\{\mu_{ik}\}$, where each element denotes the degree of membership of subject i in relation to the cluster centroid k shown in Step 5 of the Algorithm 1.

3.4. Post-soft clustering analysis

As described in the previous section, we now obtain a cluster membership vector $U_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iK})^T$ for each subject i. Considering the temporal statistical heterogeneity of the time series clinical data, i.e., two subjects with different sepsis sub-phenotypes may have opposite trends in a given time range but can still be grouped into the same cluster, we introduce an additional indicator to quantify this temporal data heterogeneity, which, in this context, pertains to similarity with the cluster centroids (sub-phenotypes). We term this indicator ABM and define it in Eq. (2) as follows:

$$ABM_i = 1 - \frac{min(d_{i1}, d_{i2}, \ldots, d_{iK})^{\frac{1}{3}}}{dist} \tag{2}$$

where $dist = max(\{d_{ik}\})$, which is computed across all subjects. The value of ABM ranges from 0 to 1, and the smaller the value, the further away it is from the cluster centroids (sub-phenotypes). We hypothesize that in datasets utilized in this study, a lower ABM value indicates increased severity of the health condition, and we further explain the hypothesis in Section 5. By combining the cluster membership vector U_i and the indicator ABM, we obtain a final representation of the soft clustering result R_i for each subject i, where we denote as $R_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{iK}, ABM_i)^T$. Intuitively, this representation captures the composition of clinical sub-phenotypes of each patient as a mixture of the three primary organ dysfunction phenotypes and the severity of the patient's health condition due to the disease.

We next use the K-Medoids clustering [37] to group all R_i to identify potential sepsis hybrid sub-phenotypes for a better classification of the soft clustering result. Note that these hybrid sub-phenotypes are combinations of the cluster centroids (sub-phenotypes) from the soft clustering results. We evaluate the quality of the clustering results by computing the mean Silhouette score [38] across all data samples and via clinical interpretation. The Silhouette score s_i for a single data sample i is computed according to Eq. (3), where a_i is the average distance of data sample i to every other samples within the assigned cluster, and b_i is the average distance of data sample i to all samples in cluster that is the closest to the assigned cluster.

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{3}$$

To optimize the post-soft clustering analysis using K-Medoids clustering, we computed the Silhouette scores for clustering results using different cluster numbers ranging from 2 to 20. Besides choosing the cluster number from the purely data-driven perspective, we also considered the medical interpretability of the data, i.e., the cluster number should be greater than 3, considering that the resulting clusters should be combinations of the clusters from the soft clustering.

3.5. Early-warning prediction

After obtaining the results of the K-Medoids clustering, we treat the resulting cluster assignment of each subject as a ground truth label of the sepsis hybrid sub-phenotype. We utilize the first 12, 24, and 48 h of the ICU data to predict the sepsis hybrid sub-phenotype of the subjects as an early-warning model. We follow the work of Lipton et al. [39] to derive statistical features from the time-series data and to compute the validation metrics. Specifically, we select seven different time windows (explained in Section 4.1) for each feature of the patient and then compute the mean, standard deviation, maximum, minimum, and skewness. Logistic regression (LR) is used for sepsis early-warning prediction due to its robustness on EHR classification tasks based on our prior work [34].

3.6. Experimental setting

We formatted all subject's data as $D \in \mathbb{R}^{N \times P \times T}$, where N is the number of subjects, P is the number of features, and T is the number of hours recorded since the ICU admission. Normalization was applied to each of the features. We set the number of clusters $K = 3$ for soft clustering, where the cluster centroids correspond to the liver, kidney, and lung dysfunction type. We selected $\eta = 2$, $\beta_1 = 10$, $\beta_2 = 0.01$ and

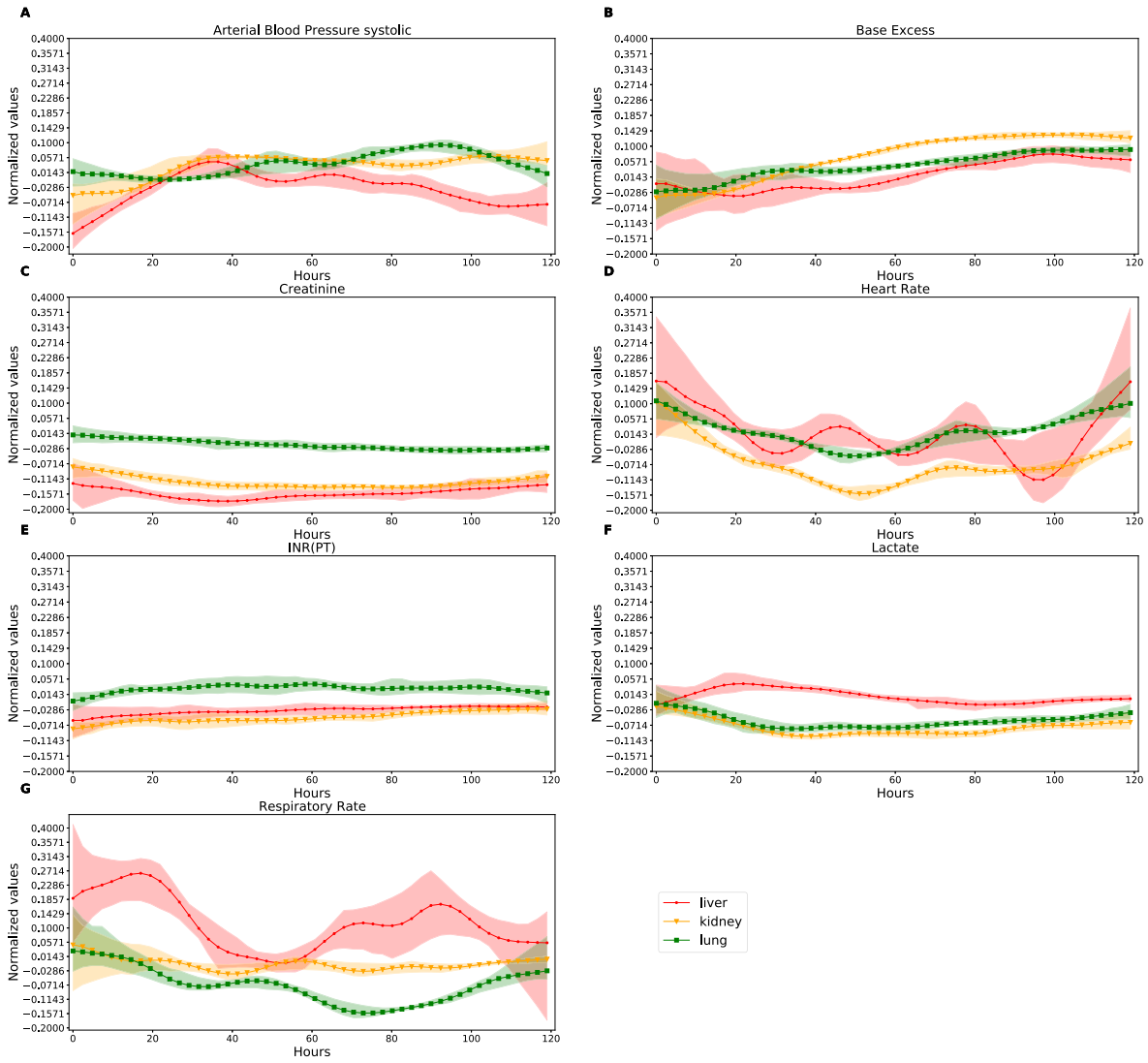

Fig. 2. Soft clustering centroids per feature after smoothing obtained from the MIMIC-IV dataset. The red centroid is initialized with the liver dysfunction type; the yellow centroid with the kidney dysfunction type; and the green centroid with the lung dysfunction type. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

conducted the iterative clustering for 200 epochs. We selected $\eta = 2$ in alignment with the convention in the literature [28], and we tested β_1 in the range of $[1 \times 10^{-2}, 10]$ and β_2 in $[1 \times 10^{-3}, 10]$. We chose the combination that yields the best Silhouette score in the post-soft clustering analysis. The SGD utilized for cluster centroids calibration was performed at a learning rate of 1×10^{-5}. The K-Medoids clustering was performed with the number of clusters equal to six. All experiments were conducted using the PyTorch library [40].

We computed features for sepsis early-warning prediction based on the work of Lipton et al. [39], where the maximum, minimum, mean, standard deviation, and skewness were computed for seven different time windows from each of the features: the entire feature sequence, the first 10%/25%/50%, and the last 10%/25%/50% of the feature sequence. Note that all features were computed from the imputed data. We applied logistic regression (LR) using the default settings for the prediction using the Scikit-learn library [41].

3.7. Evaluation metrics

We assess the clustering results by computing the average Silhouette score ranging from −1 to 1, shown in Eq. (4). The higher the score, the better the cluster quality. a_i represents the mean distance between subject i and other subjects within the cluster (intra-cluster distances), and b_i represents the mean distance between subject i and subjects from other clusters (inter-cluster distances). We compute individual Silhouette scores s_i and obtain the average value.

$$s_i = \frac{b_i - a_i}{max(a_i, b_i)} \tag{4}$$

In addition, we evaluate the sepsis hybrid sub-phenotype early-warning prediction using accuracy, precision, recall, and Area Under Precision–Recall Curve (AUPRC) following the work of Gao et al. [42].

4. Results

4.1. Evaluation on soft clustering centroids

We first present and evaluate the centroids of the three clusters obtained from the MIMIC-IV dataset shown in Fig. 2, where each of them was initialized as an organ dysfunction type and was then iteratively updated to a novel sepsis sub-phenotype after reaching convergence. In a certain sense, all results of the soft clustering algorithm can be regarded as a mixture of these three sub-phenotypes. Fig. 2 visualizes patterns of the three cluster centroids per feature. Note that

Table 1
K-Medoids cluster (hybrid sub-phenotype) centroids obtained from the MIMIC-IV dataset (left) and eICU dataset (right).

MIMIC-IV	μ_1 (liver)	μ_2 (kidney)	μ_3 (lung)	ABM	eICU	μ_1 (liver)	μ_2 (kidney)	μ_3 (lung)	ABM
Hybrid sub-phenotype 1	0.33	0.32	0.35	0.47	Hybrid sub-phenotype 1	0.33	0.34	0.33	0.36
Hybrid sub-phenotype 2	0.44	0.28	0.28	0.58	Hybrid sub-phenotype 2	0.4	0.32	0.28	0.49
Hybrid sub-phenotype 3	0.24	0.37	0.39	0.61	Hybrid sub-phenotype 3	0.24	0.35	0.41	0.51
Hybrid sub-phenotype 4	0.55	0.24	0.21	0.70	Hybrid sub-phenotype 4	0.48	0.3	0.22	0.63
Hybrid sub-phenotype 5	0.36	0.35	0.29	0.71	Hybrid sub-phenotype 5	0.31	0.35	0.34	0.63
Hybrid sub-phenotype 6	0.18	0.46	0.36	0.71	Hybrid sub-phenotype 6	0.17	0.37	0.46	0.66

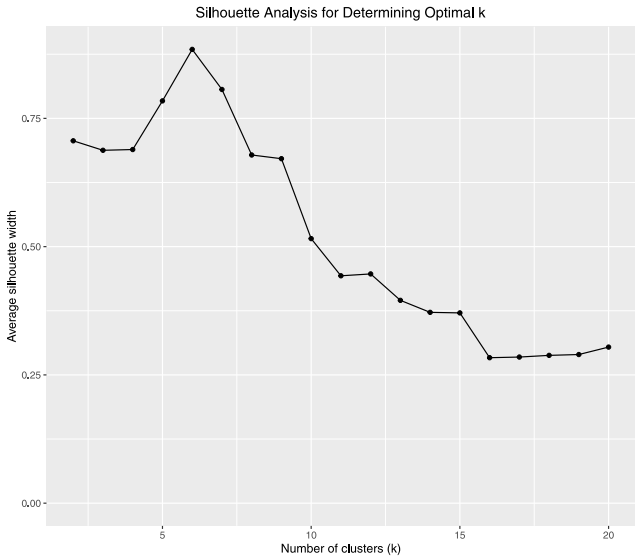

Fig. 3. Cluster number selection for K-Medoids clustering using the MIMIC-IV dataset.

Table 2
Mortality rate of each sepsis hybrid sub-phenotype group from the MIMIC-IV and the eICU datasets.

	MIMIC-IV mortality (%)	eICU mortality (%)
Hybrid sub-phenotype 1	78.71	42.73
Hybrid sub-phenotype 2	75.77	26.51
Hybrid sub-phenotype 3	71.78	12.45
Hybrid sub-phenotype 4	53.1	7.39
Hybrid sub-phenotype 5	56.15	13.74
Hybrid sub-phenotype 6	55.02	7.32

cluster centroids visualization using the eICU dataset is provided in Supplement Fig. C.13.

4.2. Post-soft clustering analysis

Fig. 3 shows the average Silhouette score computed using each cluster number. We observe that at cluster number equals 6, the cluster quality is optimal, yielding the highest Silhouette score. We thus chose 6 to be the cluster number for post-soft clustering analysis. We made the same observation using the eICU dataset from Supplement Fig. C.14.

We present in Table 1 the resulting 6 cluster centroids obtained from the MIMIC-IV and the eICU datasets, respectively, which indicate 6 potential sepsis hybrid sub-phenotypes. We rank sepsis hybrid sub-phenotypes based on the severity of patient health conditions from the most to the least severe as indicated by the ABM value. We also show the median and the interquartile range (IQR) of the clusters per feature in Figs. 4 and 5. Separate figures of feature values per hybrid sub-phenotype using the MIMIC-IV and the eICU datasets are provided in the Supplement.

In addition, we further summarize the patient outcome of each hybrid sub-phenotype group in terms of mortality rate in Table 2 to evaluate the discovered hybrid sub-phenotypes.

Table 3
Early-warning prediction results using the MIMIC-IV dataset. Precision, recall, and AUPRC are computed by the average of the "one-vs-rest" setting.

Hours	Precision	Recall	Accuracy	AUPRC
12	0.609	0.601	0.621	0.618
24	0.661	0.65	0.668	0.671
48	0.598	0.587	0.61	0.615
120	0.546	0.519	0.552	0.54

Table 4
Early-warning prediction results using the eICU dataset. Precision, recall, and AUPRC are computed by the average of the "one-vs-rest" setting.

Hours	Precision	Recall	Accuracy	AUPRC
12	0.505	0.509	0.524	0.508
24	0.569	0.565	0.576	0.565
48	0.557	0.558	0.567	0.565
120	0.531	0.531	0.54	0.547

4.3. Early-warning prediction

As mentioned in Section 3.5, we developed a sepsis hybrid sub-phenotype early-warning prediction model using the post-soft clustering results as ground truth labels. We present the results in terms of accuracy, precision, recall, and AUPRC in Tables 3 and 4 obtained from the MIMIC-IV and the eICU datasets, respectively. We compare the results with the result of using the whole 120-hour ICU data.

5. Discussion

As can be seen from Fig. 2, the three sub-phenotypes are different from each other, suggesting that the proposed semi-supervised soft clustering algorithm can generate a clear separation between the clusters. The cluster centroid in red features an elevated lactate level and a low base excess level. The cluster centroid in green exhibits elevated creatinine and INR(PT) levels. We observe that the characteristics of each of the cluster centroids do not necessarily match their patterns from the original initialization.

We observe that patients in hybrid sub-phenotype 1 obtained from both datasets exhibit the most severe health condition indicated by the lowest ABM value. The corresponding even degrees of membership to primary sub-phenotypes 1–3 ($\mu_1 - \mu_3$ in Table 1) suggest that subjects in the group may experience multiple organ failures, reflected by the lowest base excess level, the highest creatinine, INR(PT), and lactate levels compared to patients in other hybrid sub-phenotype groups.

Patients in hybrid sub-phenotypes 2 and 3 obtained from both datasets have moderate-severity health conditions. Hybrid sub-phenotype 2 subjects from both datasets have a higher degree of membership to sub-phenotype 1 (μ_1), suggesting that the subjects align more closely with liver-related characteristics, reflected by moderately high levels of lactate and INR (PT). Hybrid sub-phenotype 3 subjects from the MIMIC-IV dataset have high degrees of membership to sub-phenotypes 2 and 3 (μ_2 and μ_3), with characteristics aligned more with kidney and lung-related diseases, reflected by the moderately high

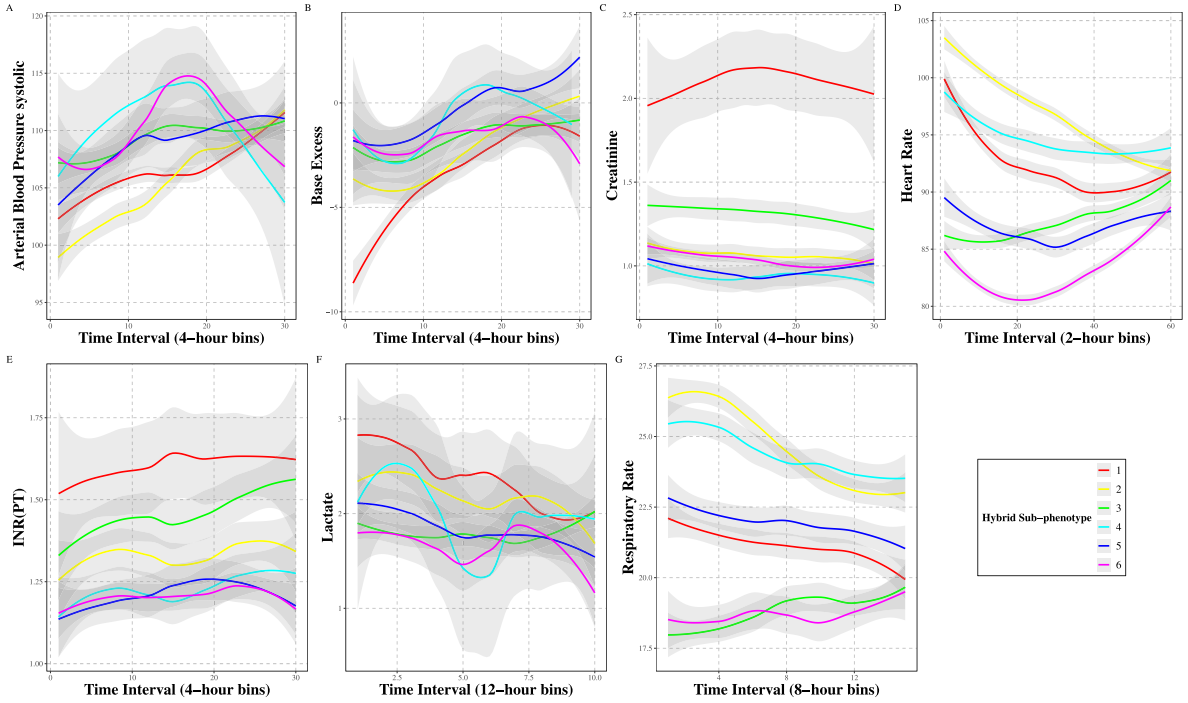

Fig. 4. Comparisons of feature values between sepsis hybrid sub-phenotypes using the MIMIC-IV dataset.

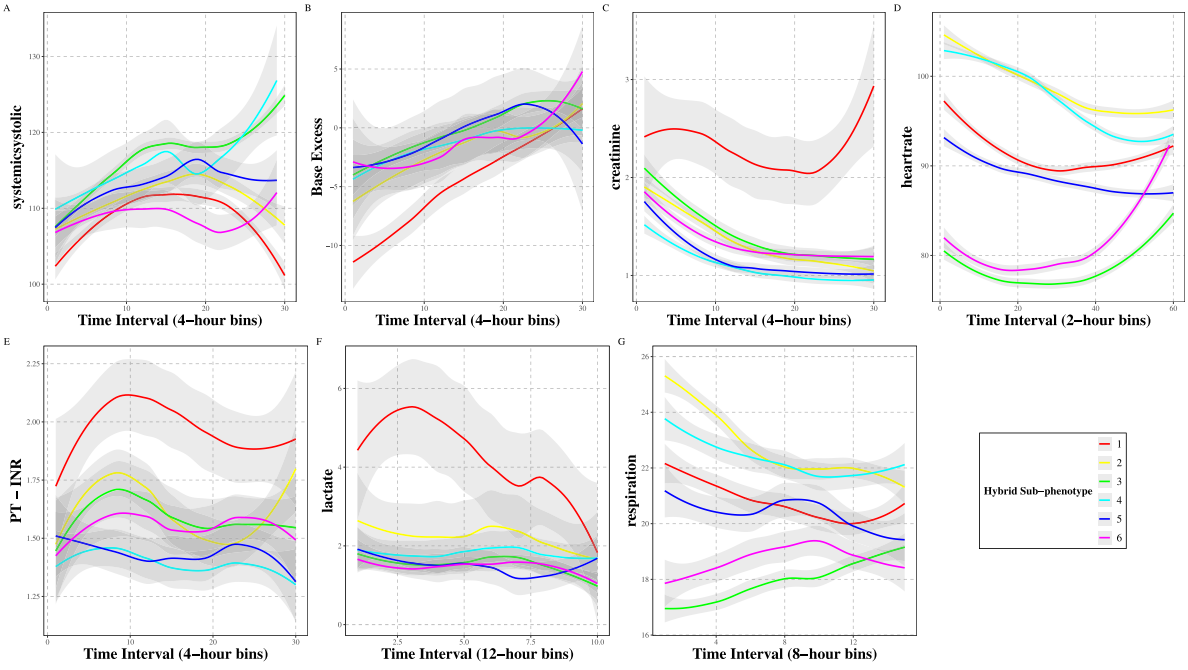

Fig. 5. Comparisons of feature values between sepsis hybrid sub-phenotypes using the eICU dataset.

creatinine level. However, we do not observe an abnormal respiratory rate. Hybrid sub-phenotype 3 subjects in the eICU dataset feature lung-related dysfunction indicated by the high degree of membership to sub-phenotype 3 (μ_3). Similarly, we do not observe an abnormal respiratory rate in the group.

Patients in hybrid sub-phenotypes 4, 5, and 6 obtained from both datasets experience a health condition of less severity. Hybrid sub-phenotype 4 subjects yield a high degree of membership to sub-phenotype 1 (μ_1), reflected by an elevated lactate level, which is consistent across both datasets.

Hybrid sub-phenotype 5 subjects from the MIMIC-IV dataset align more closely with sub-phenotypes 1 and 2 (μ_1 and μ_2), reflected by slightly elevated lactate and creatinine levels. Hybrid sub-phenotype 5 subjects from the eICU dataset do not show significant organ dysfunction, implied by the combination of even degrees of membership to all three types of primary sub-phenotypes and a higher ABM value. Hybrid sub-phenotype 6 subjects from the MIMIC-IV dataset obtain a high degree of membership to sub-phenotype 2 (μ_2), with a slightly elevated creatinine level. Hybrid sub-phenotype 6 subjects from the eICU dataset may exhibit lung-related dysfunction implied by the high

degree of membership to sub-phenotype 3 (μ_3). However, we do not observe an abnormal respiratory rate.

We find consistent patient characteristics of the hybrid sub-phenotypes 1, 2, and 4 across the MIMIC-IV and the eICU datasets. Discrepancies in hybrid sub-phenotypes 3, 5, and 6 can be attributed to the heterogeneity of patient cohorts between the MIMIC-IV and the eICU datasets. The similar findings across two distinct patient cohorts suggest that our proposed method could provide medically meaningful sepsis sub-phenotypes if given consistent datasets, which will be assessed in our future work.

We observe that hybrid sub-phenotype 1 with subjects highly associated with all three organ dysfunctions has the lowest ABM, and we notice patients in a cluster with a lower ABM, such as hybrid sub-phenotype 2, yield more abnormal feature values (e.g., creatinine and lactate levels) compared to patients in a cluster with a higher ABM, such as hybrid sub-phenotype 4. We thus hypothesize a potential association between the ABM value that is derived from a data-driven perspective with the severity of the patient's health condition based on feature values. Further exploration with other hospital data is necessary to test this hypothesis.

We observe that in the MIMIC-IV dataset, hybrid sub-phenotype 1 group exhibits the highest mortality; hybrid sub-phenotypes 2 and 3 have relatively lower mortality; hybrid sub-phenotypes 4–6 yield the lowest mortality among all the groups. This observation in mortality aligns with the ABM indicator shown in Table 1 that a lower ABM value corresponds to a higher mortality of the group. In the eICU dataset, the hybrid sub-phenotype 1 group shows the highest mortality; patients in the hybrid sub-phenotype 2 have moderate mortality; hybrid sub-phenotypes 3 and 5 exhibit low mortality; hybrid sub-phenotypes 4 and 6 yield the lowest mortality. We notice that the ABM values of hybrid sub-phenotypes 3 and 5 in the eICU dataset do not align with mortality well.

We obtain mixed results using early-hour ICU data for sepsis hybrid sub-phenotype prediction. The best prediction performance occurs when using the first 24-hour ICU data with an accuracy of 0.668 and an AUPRC of 0.671 using the MIMIC-IV dataset. Similarly, we obtain the best performance using the first 24-hour ICU data with an accuracy of 0.576 and an AUPRC of 0.565 using the eICU dataset.

The current study has some limitations. We did not include cardiovascular dysfunction as part of the clinical context of the proposed algorithm because this would also involve incorporating treatment information on vasopressor use. We plan to further evaluate the cardiovascular component as we further extend and integrate treatment information into our model. Additionally, to prove the generalizability of the proposed method, further evaluations need to be done on databases utilizing coding systems other than the ICD codes. Other improvements to consider include validation of the proposed method using private hospital data and consideration of the changes in the assigned sub-phenotypes over time. Note that phenotyping on non-ICU patients is out of the scope of this study given the wider time range and higher sparsity of the records compared to ICU data. Different approaches targeting non-ICU data will be investigated in our future work.

6. Conclusion

Sepsis sub-phenotyping is a crucial but complex area of research. To advance the classification of sepsis sub-phenotypes and incorporate temporal changes over time, we proposed a novel soft clustering algorithm that incorporates temporal and medical context using EHR data. Our results suggest the newly discovered six hybrid sub-phenotypes are medically plausible. The sepsis early-warning prediction model we created that builds upon our sub-phenotyping findings yields promising results.

Code availability

The underlying code for this study is publicly available at: https://github.com/Shiyi-J/EHR_Sepsis_Soft_Phenotyping.

CRediT authorship contribution statement

Shiyi Jiang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Xin Gai:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Miriam M. Treggiari:** Writing – review & editing, Writing – original draft, Methodology, Data curation, Conceptualization. **William W. Stead:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Methodology, Investigation, Conceptualization. **Yuankang Zhao:** Software. **C. David Page:** Writing – review & editing, Writing – original draft, Conceptualization. **Anru R. Zhang:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Anru Zhang reports financial support and administrative support were provided by Duke University. Anru Zhang reports a relationship with Duke University that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The authors declare no conflicts of interest.

Data availability

The datasets generated and/or analyzed during the current study are available in the MIMIC repository, https://mimic.mit.edu, and the eICU Collaborative Research Database, https://eicu-crd.mit.edu.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2024.104615.

References

[1] M. Singer, et al., The third international consensus definitions for sepsis and septic shock (sepsis-3), JAMA 315 (8) (2016) 801–810.
[2] R.S. Hotchkiss, L.L. Moldawer, S.M. Opal, K. Reinhart, I.R. Turnbull, J.L. Vincent, Sepsis and septic shock, Nat. Rev. Dis. Prim. 2 (2016) 16045.
[3] K.M. DeMerle, et al., Sepsis subclasses: A framework for development and interpretation, Crit. Care Med. 49 (5) (2021) 748–759.
[4] C.W. Seymour, et al., Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis, JAMA 321 (20) (2019) 2003–2017.
[5] A.R. Schertz, K.M. Lenoir, A.G. Bertoni, B.J. Levine, M. Mongraw-Chaffin, K.W. Thomas, Sepsis prediction model for determining sepsis vs SIRS, qSOFA, and SOFA, JAMA Netw. Open 6 (8) (2023) e2329729.
[6] M. Afshar, et al., Subtypes in patients with opioid misuse: A prognostic enrichment strategy using electronic health record data in hospitalized patients, PLoS One 14 (7) (2019) e0219717.
[7] M.P. Maurits, et al., A framework for employing longitudinally collected multicenter electronic health records to stratify heterogeneous patient populations on disease history, J. Am. Med. Inform. Assoc.: JAMIA 29 (5) (2022) 761–769.
[8] J. Zhao, et al., Detecting time-evolving phenotypic topics via tensor factorization on electronic health records: Cardiovascular disease case study, J. Biomed. Inform. 98 (2019) 103270.
[9] S. Mullin, et al., Longitudinal K-means approaches to clustering and analyzing EHR opioid use trajectories for clinical subtypes, J. Biomed. Inform. 122 (2021) 103889.

[10] Z. Xu, C. Mao, C. Su, H. Zhang, I. Siempos, L.K. Torres, D. Pan, Y. Luo, E.J. Schenck, F. Wang, Sepsis subphenotyping based on organ dysfunction trajectory, Crit. Care 26 (1) (2022) 197.
[11] S. Yang, P. Varghese, E. Stephenson, K. Tu, J.L. Gronsbell, Machine learning approaches for electronic health records phenotyping: A methodical review, J. Am. Med. Inform. Assoc.: JAMIA 30 (2) (2023) 367–381.
[12] T. He, et al., Trends and opportunities in computable clinical phenotyping: A scoping review, J. Biomed. Inform. 140 (2023) 104335.
[13] Y. Wang, et al., Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records, J. Biomed. Inform. (2020) 103364.
[14] Z.M. Ibrahim, H. Wu, A.A. Hamoud, L. Stappen, R.J.B. Dobson, A. Agarossi, On classifying sepsis heterogeneity in the ICU: insight using machine learning, J. Am. Med. Inform. Assoc.: JAMIA 27 (3) (2020) 437–443.
[15] W. Oh, et al., Using sequence clustering to identify clinically relevant subphenotypes in patients with COVID-19 admitted to the intensive care unit., J. Am. Med. Inform. Assoc.: JAMIA 29 (3) (2022) 489–499.
[16] Z. Xu, et al., Identifying sub-phenotypes of acute kidney injury using structured and unstructured electronic health record data with memory networks, J. Biomed. Inform. 102 (2020) 103361.
[17] T.A. Lasko, D.A. Mesa, Computational phenotype discovery via probabilistic independence, 2019, arXiv:1907.11051.
[18] J.O. Smith, C.S. Josef, Y. Xie, R. Kamaleswaran, Online critical-state detection of sepsis among ICU patients using Jensen-Shannon divergence, AMIA Annu. Symp. 2022 (2022) 982–991.
[19] H. Estiri, Z.H. Strasser, S.N. Murphy, High-throughput phenotyping with temporal sequences, J. Am. Med. Inform. Assoc. 28 (2021) 772–781.
[20] C. Lee, M. van der Schaar, Temporal phenotyping using deep predictive clustering of disease progression, in: Proceedings of the 37th International Conference on Machine Learning, Vol. 119, 2020, pp. 5767–5777.
[21] R. Xu, D.C. Wunsch, Survey of clustering algorithms, IEEE Trans. Neural Netw. 16 (3) (2005) 645–678.
[22] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, 1967, pp. 281–297.
[23] S.C. Johnson, Hierarchical clustering schemes, Psychometrika 32 (1967) 241–254.
[24] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, pp. 226–231.
[25] J.C. Bezdek, R. Ehrlich, W.E. Full, FCM: The fuzzy c-means clustering algorithm, Comput. Geosci. 10 (1984) 191–203.
[26] G. Cleuziou, An extended version of the k-means method for overlapping clustering, in: 2008 19th International Conference on Pattern Recognition, 2008, pp. 1–4.
[27] B. Zhang, M. Hsu, U. Dayal, K-Harmonic Means - A Data Clustering Algorithm, HPL-1999-124, Hewlett-Packard Laboratories, 1999.
[28] M.B. Ferraro, P. Giordani, Soft clustering, Wiley Interdiscip. Rev. Comput. Stat. 12 (1) (2020) e1480.
[29] Centers for Medicare & Medicaid Services, MS-DRG classifications and software, 2022, https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/MS-DRG-Classifications-and-Software. Accessed December, 2022.
[30] Centers for Disease Control, Classification of diseases, functioning, and disability, 2022, http://www.cdc.gov/nchs/icd.htm. Accessed October, 2022.
[31] Centers for Medicare & Medicaid Services, ICD-10-CM/PCS MS-DRG v36.0 definitions manual, 2018, https://www.cms.gov/icd10m/version36-fullcode-cms/fullcode_cms/P0326.html. Accessed December, 2022.
[32] G.P. Otto, et al., The late phase of sepsis is characterized by an increased microbiological burden and death rate, Crit. Care 15 (4) (2011) R183.
[33] P. Jain, P. Netrapalli, S. Sanghavi, Low-rank matrix completion using alternating minimization, in: Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing, 2013, pp. 665–674.
[34] S. Jiang, R. Han, K. Chakrabarty, D. Page, W.W. Stead, A.R. Zhang, Timeline registration for electronic health records, in: AMIA 2023 Informatics Summit, 2023.
[35] S. Khanmohammadi, N. Adibeig, S. Shanehbandy, An improved overlapping k-means clustering method for medical applications, Expert Syst. Appl. 67 (2017) 12–18.
[36] S. Basu, A. Banerjee, R.J. Mooney, Semi-supervised clustering by seeding, in: Proceedings of the 19th International Conference on Machine Learning, ICML-2002, 2002, pp. 19–26.
[37] H.-S. Park, C.-H. Jun, A simple and fast algorithm for K-medoids clustering, Expert Syst. Appl. 36 (2) (2009) 3336–3341.
[38] P.J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[39] Z.C. Lipton, D.C. Kale, C.P. Elkan, R.C. Wetzel, Learning to diagnose with LSTM recurrent neural networks, in: ICLR, 2016.
[40] A. Paszke, et al., Pytorch: An imperative style, high-performance deep learning library, in: Advances in Neural Information Processing Systems 32, 2019, pp. 8024–8035.
[41] F. Pedregosa, et al., Scikit-learn: Machine learning in python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[42] J. Gao, C. Xiao, L. Glass, J. Sun, Dr. Agent: Clinical predictive model via mimicked second opinions, J. Am. Med. Inform. Assoc.: JAMIA 27 (7) (2020) 1084–1091.