Received: 4 September 2021 | Revised: 22 November 2021 | Accepted: 30 December 2021

DOI: 10.1002/emp2.12660

JACEP OPEN
JOURNAL OF THE AMERICAN COLLEGE OF EMERGENCY PHYSICIANS OPEN
A GLOBAL JOURNAL OF EMERGENCY MEDICINE
WILEY

ORIGINAL RESEARCH

Pediatrics

Pediatric sepsis phenotypes for enhanced therapeutics: An application of clustering to electronic health records

Ioannis Koutroulis MD, PhD[1] | Tom Velez PhD[2] | Tony Wang PhD[3] | Seife Yohannes MD[4] | Jessica E. Galarraga MD, MPH[4] | Joseph A. Morales[2] | Robert J. Freishtat MD, MPH[1] | James M. Chamberlain MD[1]

[1] Emergency Medicine, Children's National Hospital/George Washington University School of Medicine and Health Sciences, Washington, District of Columbia, USA

[2] Computer Technology Associates, Cardiff, California, USA

[3] Imedacs, Ann Arbor, Michigan, USA

[4] MedStar Health Research Institute, Hyattsville, Maryland, USA

Correspondence
Ioannis Koutroulis, Emergency Medicine, Children's National Hospital/George Washington University School of Medicine and Health Sciences, Washington, DC, USA.
Email: ikoutrouli@childrensnational.org

Funding information
NIH National Institute of General Medical Sciences, Grant/Award Number: 1R43GM122154-01

Abstract

Objective: The heterogeneity of pediatric sepsis patients suggests the potential benefits of clustering analytics to derive phenotypes with distinct host response patterns that may help guide personalized therapeutics. We evaluate the relative performance of latent class analysis (LCA) and K-means, 2 commonly used clustering methods toward the derivation of clinically useful pediatric sepsis phenotypes.

Methods: Data were extracted from anonymized medical records of 6446 pediatric patients that presented to 1 of 6 emergency departments (EDs) between 2013 and 2018 and were thereafter admitted. Using International Classification of Diseases (ICD)-9 and ICD-10 discharge codes, 151 patients were identified with a sepsis continuum diagnosis that included septicemia, sepsis, severe sepsis, and septic shock. Using feature sets used in related clustering studies, LCA and K-means algorithms were used to derive 4 distinct phenotypic pediatric sepsis segmentations. Each segmentation was evaluated for phenotypic homogeneity, separation, and clinical use.

Results: Using the 2 feature sets, LCA clustering resulted in 2 similar segmentations of 4 clinically distinct phenotypes, while K-means clustering resulted in segmentations of 3 and 4 phenotypes. All 4 segmentations identified at least 1 high severity phenotype, but LCA-identified phenotypes reflected superior stratification, high entropy approaching 1 (eg, 0.994) indicating excellent separation between estimated phenotypes, and differential treatment/treatment response, and outcomes that were nonrandomly distributed across phenotypes ($P < 0.001$).

Conclusion: Compared to K-means, which is commonly used in clustering studies, LCA appears to be a more robust, clinically useful statistical tool in analyzing a heterogeneous pediatric sepsis cohort toward informing targeted therapies. Additional prospective studies are needed to validate clinical utility of predictive models that target derived pediatric sepsis phenotypes in emergency department settings.

Supervising Editor: Chadd Kraus, DO, DrPH

KEYWORDS
K-means, LCA, phenotypes, sepsis

1 | INTRODUCTION

1.1 | Background

In the absence of routinely available biomarkers, recent large cohort studies indicate that clustering routinely available patient electronic health record (EHR) data can identify clinically useful phenotypes toward personalized care.[1] The most common clustering techniques used in medicine[2] are mixture models exemplified by latent class analysis (LCA)[3] and latent profile analysis (LPA),[4] and K-means.[5] In domains relevant to our study, LCA has been used to detect phenotypes in septic shock[6] and ARDS,[7–9] LPA in adult sepsis,[10] whereas K-means has been used in critically ill pediatric patients,[11] adult sepsis,[12,13] and septicemia.[14] In many of these studies, derived phenotypes were considered useful as a basis for targeted therapies. A comparable derivation of EHR data-based pediatric sepsis phenotypes has not been reported. Unlike K-means that uses a distance metric to identify clusters, mixture model techniques estimate the probability that a given patient belongs to each of the different latent classes and are therefore considered more statistically robust, accurate methods of clustering.[15–17] The clinical impact of this reported difference in clustering methods has not been studied.

1.2 | Importance

Finding clusters of similar patients whose cluster membership provides information likely pertinent to prognosis and response to therapy is an important goal in sepsis management. Although compliance with pediatric sepsis protocols or "bundles" has been associated with improved outcomes,[18–20] it has been hypothesized that sepsis heterogeneity may explain why most therapeutic interventions have not improved overall sepsis survival.[21] Homogenized treatment given to a highly heterogeneous patient population that fall under the wide syndromic umbrella of sepsis may offer a minimum standard of care although at the potential cost of compromising outcomes in an individual patient.[22] For example, although intravenous fluid boluses remain a cornerstone of the resuscitation of children with septic shock, an increasing number of publications have highlighted the increased morbidity and mortality associated with aggressive fluid administration[23] and the benefits of early vasopressors.[24] The basic concept of precision medicine is the early identification of specific phenotypes of patients that will respond to phenotype-targeted personalized treatment (eg, type/amount/timing of fluids, vasopressors), or other treatments.[25] Tools to provide this guidance not used in the practice of pediatric sepsis, and consequently care across the highly heterogeneous pediatric sepsis population, remains homogenized.[22] This situation is exacerbated by recent quality improvement studies that redefined pediatric sepsis to include not only those with infection-related organ dysfunction but also those with infections that were treated for sepsis with organ dysfunction potentially averted.[26]

1.3 | Goals of this investigation

Ideally, from a clinical perspective, clustering of the EHR data associated with a heterogeneous population would identify well-separated,[17,27] biologically plausible, reproducible homogeneous subgroups with distinct severities (risk stratification) and treatment responses that would prognostically inform effective treatment. Both LCA and K-means have been used in clustering studies cited above as if these techniques are equivalent. However, in a study comparing the accuracy of LCA and K-means, clustering in correctly identifying classes where true class-membership was known but concealed during analyses found that misclassification rate was approximately 4 times higher using K-means clustering than LCA.[28] We performed this study to derive and compare the potential clinical use of phenotypes obtained by LCA and K-means in a cohort of pediatric sepsis patients. Given that a comparable pediatric sepsis phenotyping effort has not been reported, we used 2 differing, yet relevant feature sets for our study: the 22 clinical Pediatric Risk of Mortality (PRISM)-based[29] features used by Williams et al[11] to derive 10 phenotypes of critically ill pediatric patients in intensive care, and the 29 clinical features (age-adjusted for our study) used by Seymour et al[12] to derive 4 phenotypes of sepsis patients meeting the most recent adult sepsis definitions. Our goal was to compare the robustness of LCA and K-means to varying feature sets of physiological variables, the separation quality of identified phenotypes considering both distance and entropy metrics, and most importantly, the clinical use of derived phenotypes toward informed outcomes and personalized therapies.

2 | METHODS

2.1 | Study design and setting

This is an institutional review board-approved observational descriptive retrospective[30] study of the EHRs of 6446 non-neonatal pediatric patients presenting to 1 of 6 tertiary adult/pediatric care admitting facilities operated by a single institution that occurred from 2013–2018. The primary data source was the MedStar Washington Hospital Center, a large academic medical center in Washington, DC, with additional data extracted from 5 additional MedStar hospitals located in the DC-Baltimore metropolitan area. MedStar has the second-largest pediatric liver transplant program in the United States.

2.2 | Selection of participants

In this cohort of 6446 hospitalized patients, 151 were identified that met Improving Pediatric Sepsis Outcomes (IPSO) sepsis criteria defined by EHR evidence of suspected infection and sepsis treatment or organ dysfunction. Identification included hospitalized pediatric patients with International Classification of Diseases (ICD)-9 and ICD-10 discharge codes as listed in Supporting Information Table S1 representing a broad sepsis continuum diagnosis that included severe infections, septicemia, sepsis, severe sepsis, and septic shock. Unlike supervised machine-learning that relies on accurate positive/negative case labeling, unsupervised clustering algorithms seeking subclasses of a condition are trained using training data associated with only positive cases. Although sepsis is routinely under-recognized[31] and under-reported resulting in poor sensitivity, sepsis ICD diagnostic codes have been shown to have high specificity[32] and infection codes (eg, septicemia) effective in detecting patients with infections associated with pediatric sepsis.[33]

2.3 | Interventions

Participating institution extracted demographic and physiological data of admitted pediatric patients from the common EHR system (Cerner) used at the participating hospitals of pediatric patients that were admitted over the 5-year period at any 1 of their 6 hospitals and identified those with a sepsis diagnosis. The institution transferred de-identified data to a central data warehouse for analysis.

2.3.1 | EHR features used in clustering analysis

The EMR recorded clinical observational data of the 151 identified sepsis patients used to derive features used in LCA and K-means modeling (independent variables) did not contain information about diagnosis, treatments, or outcomes (dependent variables). We studied 2 distinctive sets of independent variables used to derive pediatric sepsis phenotype-defining features:

1. Twenty-two variables based on PRISM score[34] used by Williams et al[11] to cluster pediatric patients receiving critical care. The variables used by Williams et al[11] included inflammatory (temperature, white blood cell count [WBC]), pulmonary (respiratory rate [RR], oxygen saturation [SpO_2], partial pressure of oxygen [PaO_2], partial pressure of carbon dioxide [$PaCO_2$], fraction of inspired oxygen [FiO_2]), renal (creatinine [Cr]), hepatic (total bilirubin), cardiac/hemodynamic (heart rate [HR], bicarbonate, mean arterial pressure [MAP]), hematologic (platelets, partial thromboplastin time [PTT], red blood cell count [RBC]), central nervous system (CNS), Glasgow Coma Scale (GCS), and acid-base and electrolytes (pH, potassium, sodium, calcium, chloride).
2. Twenty-nine variables used by Seymour et al[35] based on their association with sepsis onset or outcome, their incorporation in conceptual models of sepsis pathophysiology and host tolerance, and their availability in the electronic health record at hospital presentation to cluster adult patients meeting Sepsis-3 criteria. The variables used by Seymour et al[35] were age-adjusted as appropriate[11] and included: age, Elixhauser comorbidity index, inflammatory (temperature, WBC, bands, erythrocyte sedimentation rate [ESR], and C-reactive protein [CRP]), pulmonary (RR, SaO2, PaO_2,), renal (blood urea nitrogen [BUN], Cr), hepatic (aspartate transaminase [AST], alanine transaminase [ALT], total bilirubin), cardiac/hemodynamic (HR, bicarbonate, systolic blood pressure [SBP], lactate, troponin), hematologic (platelets, international normalized ratio [INR], hemoglobin [Hgb]), CNS (GCS), and serum levels of glucose, sodium, chloride, and albumin.

The Bottom Line

Applying statistical clustering methods, such as latent class analysis (LCA) and K-means, can help to derive clinically relevant models for pediatric sepsis phenotypes to aid in diagnostic and therapeutic approaches in the emergency department (ED). Using data from nearly 6500 pediatric patients in 6 EDs over 5 years, LCA appears to be a more robust, clinically useful tool in analyzing a heterogenous pediatric sepsis cohort toward informing targeted therapies.

EMR data is typically noisy and may include erroneous data, redundancies, and semantically equivalent or similar observations measured by more than 1 technique (eg, invasive vs non-invasive blood pressure, skin vs core temperature, etc). We removed clearly erroneous data (eg, non-numeric or physiologically impossible outliers), semantically harmonized similar measurements and, based on established associations between lab values and organ dysfunction criteria,[36] used the most abnormal minimum or maximum value of each variable over the entire period of hospital stay for the LCA and K-means analysis as shown in Supporting Information Tables S2 and S3.

In both Williams et al[11] and Seymour et al[35] features, significant "missingness" was observed. This was expected given the diversity of sepsis severity (eg, from disseminated infections treated in wards to those with MODS treated in PICUs) By manual chart review, we validated that specific "missing values" (eg, arterial blood gasses, missing metabolic panel labs in certain patients) were not "at random" and were primarily associated with tests not indicated for the sepsis type and severity (eg, those treated for infection with no organ dysfunction). Consequently, missing observations in these patients were replaced with age-adjusted normal values.[11]

For age-dependent features including heart rate, respiratory rate, blood pressure, and creatinine, we used predetermined age cohorts: 0–1, 2–5, 6–8, 9–12, and 13–18 years (age at admission) and z normalized using the data to define age group means and standard deviations and transformed back to the measurements of the 13–18 year age cohort, to facilitate clinical interpretation of the results. Data were

pre-processed with Box-Cox transformation (for model-based clustering) and then z normalization before clustering analysis.

2.4 | LCA model derivation

For LCA, to select a model fitting the data best, a series of probabilistic Gaussian mixture models[37] using clinical features with different number of components are fitted,[38] and the Bayesian Information Criteria (BIC) is used for model selection.[39] Latent class model estimation is based on full-information maximum likelihood methods. LCA models were developed using the R package "mclust 5."[39]

2.5 | K-means model derivation

Consensus K-means clustering is based on calculated Gower distance matrix,[40] which takes mixed continuous variables and categorical variables into consideration. Proportion of ambiguous clustering (PAC), defined as the fraction of sample pairs with consensus index values falling in the intermediate sub-interval, is used for identifying optimal number of clusters.[41] Once optimal number of clusters is determined, the final cluster is obtained based on 100 resamples.[41] K-means models were developed using "ConsensusClusterPlus," a class discovery R package.[42]

The 4 derived phenotypic models: LCA/Williams, LCA/Seymour, K-means/Williams and K-means/Seymour were evaluated for sample size adequacy, clustering internal validity (distance-based compactness, connectedness, and separation), entropy as a measure of LCA class separation, robustness to features, phenotype homogeneity, clinical use (distinct diagnosis, host-response to treatment, length-of-stay, and mortality).

2.6 | Statistics/evaluating adequacy of sample size

Given the relatively small number of sepsis cases in this study population (N = 151) Monte Carlo simulations were used to examine the adequacy of power in this sample size to detect the "true" number of latent classes and consensus clusters.[17,43] Based on estimated mean and covariance structures from Williams and Seymour features, data with different sample sizes are generated based on multivariate normal distribution. LCA and Consensus Clustering analysis are then applied on simulated data. The objective of the simulation study is to examine the relationship between sample sizes and the power to detect the "true" number of latent classes. Power is calculated as the proportion of the number of simulations which identifies correct number of components out of 1000 simulations.

2.7 | Statistics/evaluating clustering internal validity

Clustering validity was evaluated using internal distance-based measures that reflect the compactness, connectedness, and separation of the cluster partitions.[44] Connectedness relates to what extent observations are placed in the same cluster as their nearest neighbors in the data space, and is here measured by the connectivity metric, which has a value between zero and ∞ and should be minimized.[45] Compactness assesses cluster homogeneity, usually by looking at the intra-cluster variance, while separation quantifies the degree of separation between clusters (usually by measuring the distance between cluster centroids). Because compactness and separation demonstrate opposing trends, these measures are commonly combined into single scores: the Dunn index[46] and silhouette width,[47] representing non-linear combinations of the compactness and separation measures. The Dunn index is the ratio of the smallest distance between observations not in the same cluster to the largest intra-cluster distance, has a value between zero and ∞, and should be maximized.[45] The silhouette width is the average of each observation's silhouette value. The silhouette value measures the degree of confidence in the clustering assignment of an observation, with well-clustered observations having values near 1 and poorly clustered observations having values near −1. The silhouette width has values in the interval [−1, 1], and should be maximized.[45] The 3 internal measures connectivity, Dunn index, and silhouette width were calculated using the R package "clValid".[45]

2.8 | Statistics/evaluating LCA class separation

To evaluate how well the latent classes are identified,[17] entropy was calculated. We used the formula provided by Asparouhov[27]:

$$E = 1 + \frac{1}{N\log(k)}\left(\sum_{i=1}^{N}\sum_{k=1}^{K} P(C=k|U_i)\log(p(C=k|U_i))\right),$$

where C is the latent variable, K is the number of classes, N is the sample size, U_i is the vector of all latent class indicator variables, and the probabilities $P(C=k|U_i)$ are computed from the estimated model.

The larger the entropy is, the clearer the latent class identification is. The entropy value is between 0 and 1. Entropy with values approaching 1 indicate clear separation of the classes.

3 | RESULTS

3.1 | LCA-based phenotypes

3.2 | LCA phenotypes using Williams features: LCA/Williams

The Williams features (Supporting Information Table S2) were used to derive a series of Gaussian models with different numbers of partitions. As shown in Supporting Information Figure S1, the Bayesian information criterion (BIC) suggests that a 4-component VEI clustering model (ie, varying volume, equal shape, and axis parallel orientation[48]) fits the data best. Simple statistics of the partitioned phenotype-defining independent features and dependent/outcome

TABLE 1 LCA/Williams phenotype clinical characteristics

Phenotype	Phenotype 1 critical severity	Phenotype 2 low severity	Phenotype 3 moderate severity	Phenotype 4 high severity
Key clinical characteristics	MODS: neurological dysfunction, renal dysfunction, thrombocytopenia, tachycardia, tachypnea and severe hypoxia	No OD: no abnormalities in labs/vitals other than fever and elevated WBC (SIRS)	No life-threatening OD: mild tachypnea, elevated LFTs	OD: overt liver dysfunction with hypoxia, mild hyponatremia
Clinician diagnosed severity	Critical: most number of sepsis/shock diagnoses (86%)	Least severe: (adjusting for a carry forward shock code), no patients had a severe sepsis or shock diagnosis	Moderate severity: includes patients clinically diagnosed with severe sepsis/septic shock (11%)	High severity: includes significant number of severe sepsis/shock cases (22%)
Treatment implications	Patients in this type will likely require critical care: vasopressors, MV and develop hyperchloremic metabolic acidosis. Prolonged LOS. Mortality	Patients in this type will likely not require critical care, vasopressors or MV. Short LOS.	Patients in this type will likely require critical care, will not require vasopressors or MV. Short LOS.	Patients in this type will likely require critical care, vasopressors, non-invasive MV. Prolonged LOS.

Abbreviation: GCS, Glasgow Coma Scale; LCA, latent class analysis; LFT, liver function test; LOS, length of stay; MODS, multiple organ dysfunction syndrome; MV, mechanical ventilation; OD, organ dysfunction; SIRS, systemic inflammatory response syndrome; WBC, while blood cell count.

variables across the 4 LCA/Williams phenotypes shown in Supporting Information Table S4. As shown in this table, the numbers of patients assigned into risk-stratified phenotypes are 23 in phenotype 1 (best described as the "critical severity" phenotype), 40 in phenotype 2 (the "low severity" phenotype), 38 in phenotype 3 (the "moderate severity" phenotype), and 50 in phenotype 4 (the "high severity" phenotype).

The key clinical features, risk stratification, and treatment implications associated with these 4 phenotypes are summarized in Table 1, and boxplots of the distribution of Williams clinical features across phenotypes are shown in Figure 1. As shown in this boxplot, in phenotype 1 there was multi-organ dysfunction. More specifically, the pH, PaO_2, SpO_2 and bicarbonate were low and respiratory rates high, indicating respiratory failure and the need for mechanical ventilation. Additionally, MAP and GCS medians suggested cardiac and CNS dysfunction.

In general, LCA/Williams phenotype 1 is characterized as MODS with the highest proportion of patients using vasopressors, requiring invasive mechanical ventilation (MV), and developing hyperchloremic metabolic acidosis. LCA/Williams phenotype 4 patients had the highest average of total bilirubin, indicating liver dysfunction, with mild hyponatremia. Notably, 2 phenotypes representing approximately 50% of this population did not have life-threatening OD but received sepsis treatment and a sepsis diagnosis (consistent with IPSO sepsis criteria[26]). LCA/Williams phenotype 2 characterized as infection-induced SIRS that would respond to fluids/antibiotics and phenotype 3 characterized as moderate severity OD that included infected high risk immunocompromised (liver transplant) patients requiring critical care that were, in some cases given a severe sepsis/septic shock diagnosis despite lack of overt critical OD requiring vasopressors or MV (exemplifying the discord between consensus criteria and physician diagnoses). It was noted (Supporting Information Table S4) that among the traditional sepsis biomarkers,[49] lactate and bands were significantly distributed across phenotypes compared to CRP and ESR. The high ALT/AST biomarkers in phenotype I were consistent with the fact that the cases were obtained by a major pediatric transplant center and many of the septic patients were susceptible to liver damage. The pH, PaO_2, SPO_2, and bicarbonate were lower and respiratory rates higher in phenotype 1. This explains the higher proportion of acidosis and respiratory failure with the need for MV. Similarly MAP and GCS medians were lower in phenotype 1 compared to the other phenotypes, suggesting that there were many patients with cardiac and CNS dysfunction.

3.2.1 | LCA phenotypes using Seymour features: LCA/Seymour

As above, the Seymour features were used to derive a series of Gaussian models with different numbers of partitions. As shown in Supporting Information Figure S2, BIC criteria also resulted in a 4-phenotype model (Supporting Information Table S5).

LCA/Seymour phenotype 1 patients were characterized by tachycardia, tachypnea, and hypoxia. This phenotype is also defined by bandemia, electrolyte abnormalities such as hyperchloremia and hyponatremia, hypoalbuminemia, and the need for vasopressors. Multiple patients in this phenotype were given severe sepsis/shock diagnoses. From a risk stratification perspective, this LCA Seymour type 1 was best matched with the LCA/Williams phenotype 4 cluster (high severity). LCA/Seymour phenotype 2 patients showed no major abnormalities in laboratory or physiologic findings other than fever and mildly elevated WBC, indicating SIRS but included some severe sepsis/shock cases, and were best matched with LCA/Williams phenotype 3 (moderate severity), whereas LCA/Seymour phenotype 3 was observed having almost no abnormal values, best matched with LCA/Williams phenotype 2 (low severity). Phenotype 4 was comprised of the sickest

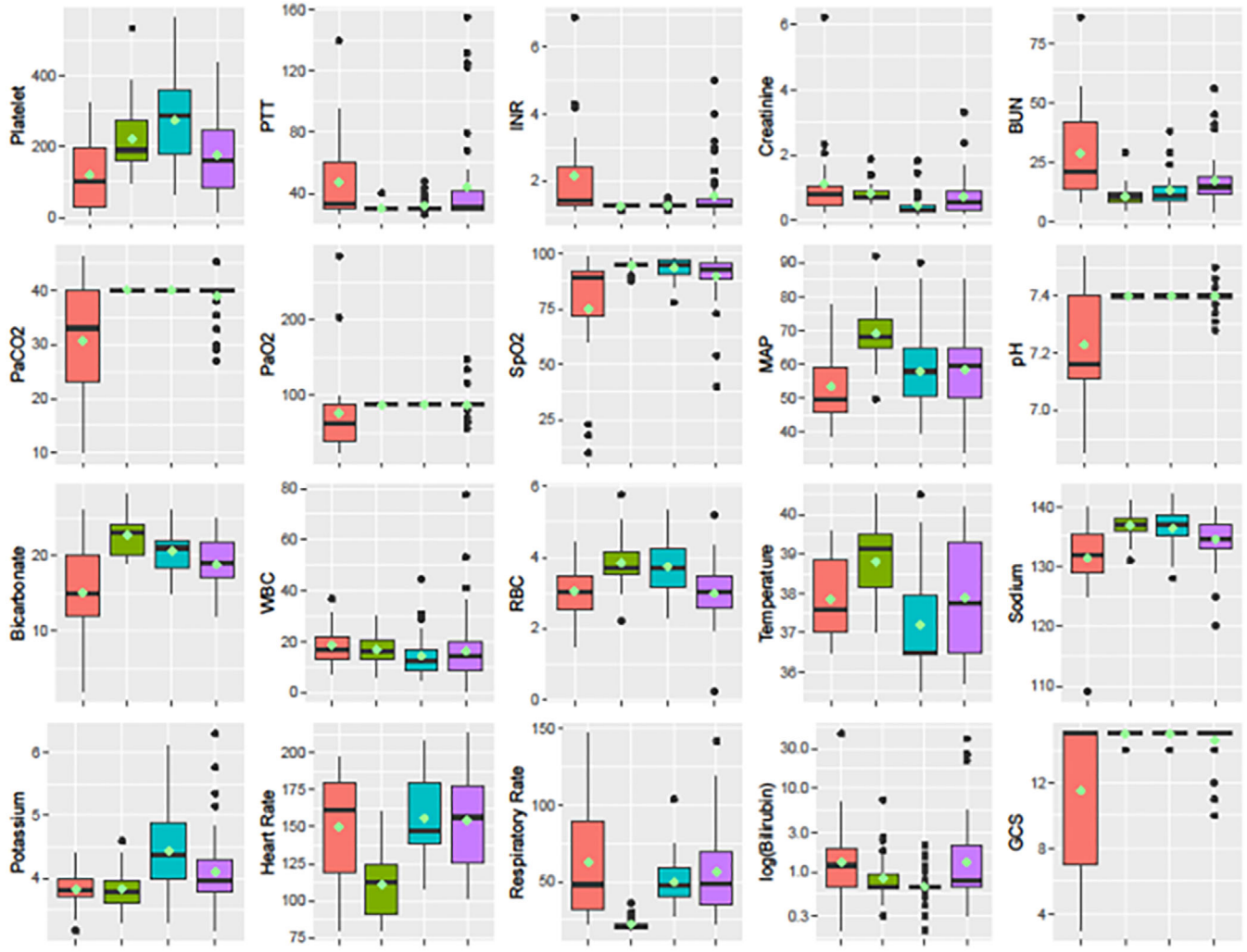

FIGURE 1 Distribution of Williams features across phenotypes. Red, phenotype 1 (critical severity); green, phenotype 2 (least severe); blue, phenotype 3 (moderate severity) and purple, phenotype 4 (high severity)

TABLE 2 Metrics between Williams and Seymour segmentations

Metric[a]	Phenotype 1	Phenotype 2	Phenotype 3	Phenotype 4
Sensitivity	0.591	0.842	0.750	0.493
Specificity	0.922	0.929	0.843	0.798
PPV	0.565	0.800	0.474	0.660
NPV	0.930	0.946	0.947	0.663
Balanced accuracy	0.757	0.886	0.796	0.645
Overall kappa		0.497		

Abbreviations: NPV, negative predictive value; PPV, positive predictive value.
[a]Metrics are calculated with the method of 1 versus all other phenotypes.

patients in the cohort bests matched with LCA/Williams phenotype 1 (critical severity). Table 2 shows class-specific performance metrics when considering using 1 phenotyping result to predict another. The κ coefficient of overall agreement is 0.497 with P value < 0.001, indicating moderate reproducibility.

It is worth mentioning the distribution of race across LCA/Williams phenotypes was statistically significant ($P = 0.02$ as shown in Supporting Information Table S4 with distribution detailed in Supporting Information Table S4A) revealing that patients belonging to non-Hispanic Black or other non-Caucasian race were more likely to be assigned to

higher severity phenotypes (phenotypes 1 and 4). Our LCA/Seymour analysis did not yield similar statistically significant results ($P = 0.25$ as shown in Supporting Information Table S5).

3.3 | K-means-based phenotypes

3.3.1 | K-means/Williams

Our K-means/Williams clustering analysis is presented in Supporting Information Figure S3, which resulted in a 3-phenotype model with features and dependent variables distributed across phenotypes as shown in Supporting Information Table S6. Among the 3 K-means/Williams phenotypes, phenotype 3 patients were the sickest, whereas phenotypes 1 and 2 were similar in severity and did not seem to have distinct characteristics that would make them clinically useful.

3.3.2 | K-means/Seymour

The comparable K-means analysis using Seymour features, as shown in Supporting Information Figure S4, resulted in a 4-phenotype model (Supporting Information Table S7). Phenotype 2 patients were characterized by MODS, with vital sign changes and electrolyte and liver function test abnormalities. These were the sickest patients of the cohort, with hyperglycemia, a bleeding diathesis, anemia, lactic acidosis, and bandemia. We were not able to identify distinguishing characteristics in patients of phenotypes 1, 3, and 4, but they included patients with a lower disease severity compared to phenotype 2. Although phenotypes 1 and 3 had very analogous results and vital signs with patients that had a milder disease than phenotype 2, they still had some patients that required vasopressors, which means that intravenous (IV) fluid boluses were not able to improve the observed hypotension. Phenotype 1 also included a significant number of patients in septic shock (21.2%), which was similar to phenotype 2, indicating that this model was not able to separate potentially sick patients from those that would have a milder course.

3.4 | Validation

3.4.1 | Adequacy of sample size

When $N < 300$, standard power calculations should be performed to determine the sample size needed to detect significant interclass differences.[17] Figures 2 and 3 show the relationship between sample size and power to detect correct number of components in simulated data. They indicate that both LCA and consensus clustering analysis achieves at least 80% power with the sample size of 150.

Clustering internal validity

To compare clustering internal validity of LCA and K-means methods independent of feature sets used, validity measures were evaluated on the union of Williams and Seymour features and shown in Table 3. Using these measures, compared to LCA, consensus K-means clustering on Seymour features shows minimized values in connectivity and maximum value in silhouette width indicating for this heterogeneous pediatric dataset, K-means demonstrates superior internal validity metrics. This result is not surprising given that internal distance-based measures such as compactness, connectedness, and separation are all based on distance measures optimized by the K-means algorithm.

TABLE 3 Calculated internal validity distance-based metrics

	LCA		K-means	
Measure	Williams	Seymour	Williams	Seymour
Connectivity	197.8	159.3	134.5	75.7
Dunn	0.003	0.003	0.014	0.081
Silhouette	−0.075	−0.048	0.146	0.191

Abbreviation: LCA, latent class analysis.

3.5 | Entropy index of class separation

Beyond distance-based cluster validation, we calculated entropy, a measure of class separation,[17] that can be informative of how well the clusters differentiate by measuring how distinctly each patient's estimated phenotype is. Entropy ranges from 0 to 1 with values close to 1 indicating a high probability of patients being in just 1 class. Given the fundamental differences in posterior probabilities-based (LCA) versus distance-metric-based (K-means) clustering, a comparative entropy metric across these 2 clustering techniques was not available.[50] However, for the LCA methods, we found that LCA/William's entropy = 0.994 and LCA/Seymour's entropy = 0.997, indicating both LCA models exhibited good phenotypic separation.

3.6 | Limitations

3.6.1 | Features and missingness

We did not attempt to optimize features to the specific 151-case IPSO sepsis population used in this study and instead chose features already validated in the literature to derive consensus-criteria defined sepsis phenotypes. Although the features used demonstrated statistically significant ($P < 0.001$) clinical distinctions across LCA versus K-means derived phenotypes, our results may have been biased by high levels of missingness. Additionally our features did not reflect clinician gestalt, known to be especially effective in the management of low and high risk cases,[51] as additional evidence beyond recorded therapies and diagnostic tests. Finally, although restricting the features used for phenotyping modeling to those available in routine clinical practice is reasonable, studies indicate that additional evidence such as advanced biomarkers driven by genomics may contribute in pediatric

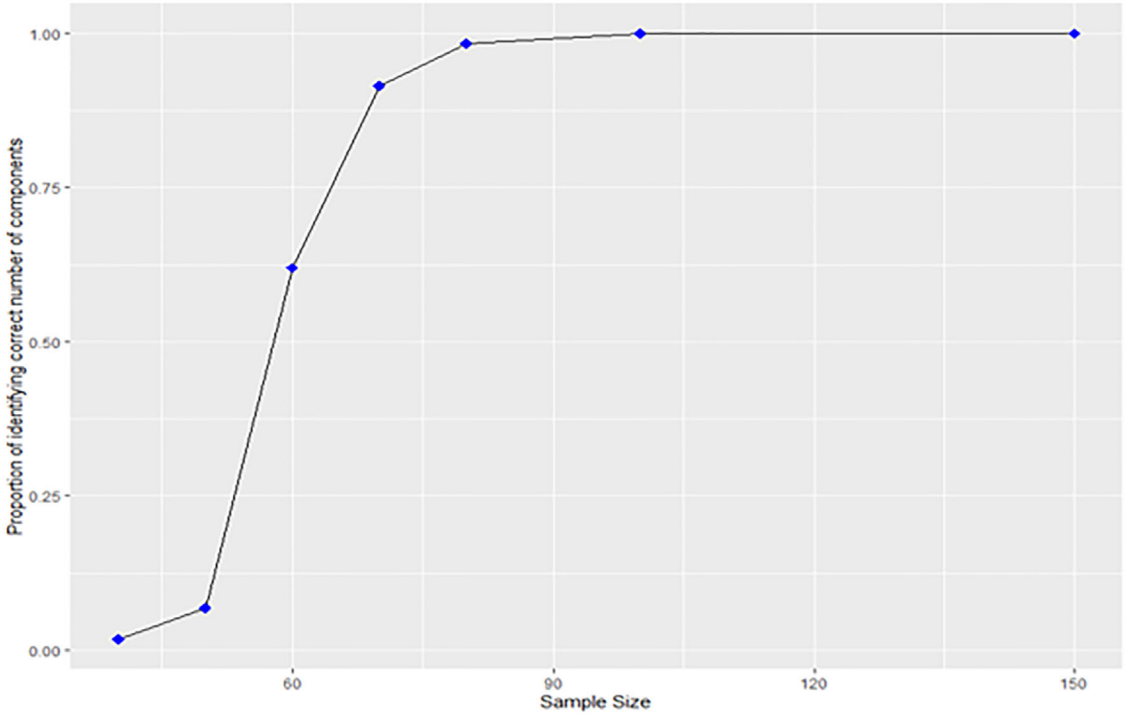

FIGURE 2 Relationship between power and sample size for LCA

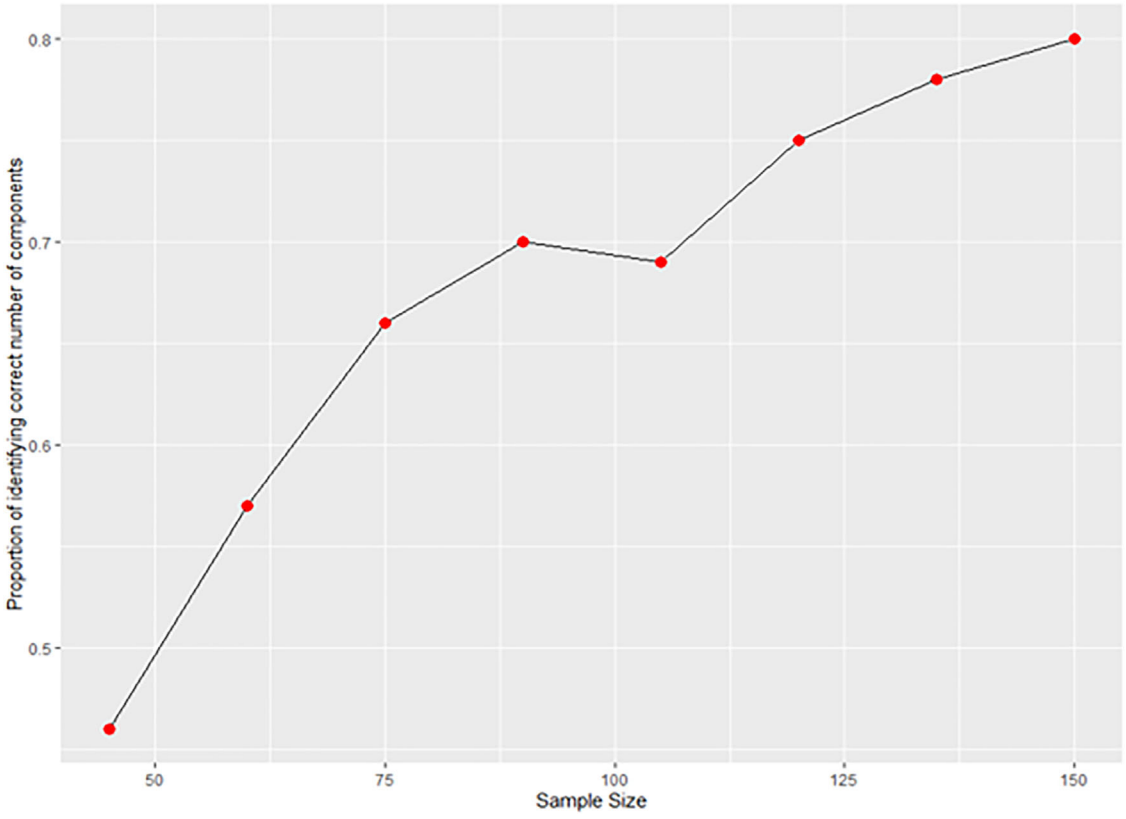

FIGURE 3 Relationship between power and sample size for K-means

sepsis phenotyping.[52,53] A more optimized pediatric set of features should be pursued in future studies.

3.6.2 | Generalizability

Given the strong institution specific bias in population characteristics used to derive clusters (eg, a significant number of liver transplant cases in our studied population), our derived phenotypes are not likely "generalizable" and applicable to other institutions serving distinctly different populations. There is a growing recognition that the narrow focus on generalizability of machine learning-based models should be replaced with the derivation and routine maintenance of institution-specific (ie, continuously learning) models that are useful at bedside.[54] In this article, we compared the 2 different clustering methods to examine whether more commonly used K-means would perform similarly to LCA. Our study highlights the superiority of LCA that, despite the need for higher computing power, can be the basis for personalized medicine in other population studied.

3.6.3 | Clinical use

Our study was based on phenotypes derived using a hospitalized population. Although the use of hospitalized children allowed us to include patients with sepsis of different severity that in the end will help identify patients early on and, whereas in the ED, the use of EHR-data derived phenotypes in clinical settings, such as EDs, will require derivation of predictive models using early data captured in the ED that target phenotype membership of presenting patients and prospective studies demonstrating use in guiding personalized treatment.

4 | DISCUSSION

Although homogenizing treatment given to a highly heterogeneous patient population that fall under the wide syndromic umbrella of pediatric sepsis may offer a minimum standard of care, it is unlikely to offer optimal care to an individual patient, potentially compromising outcomes.[22,55] This is one of the first studies to report on routine EHR-data driven phenotypes associated with seriously infected hospitalized children that, despite recognized heterogeneity, currently typically receive protocolized treatment.[56]

There is no consensus on the best distance function, clustering method, or feature selection to be used with EHR data to detect subgroups of a heterogeneous patient population that can inform personalized treatment.[44] From a feature selection perspective, when comparing LCA phenotypes described in Supporting Information Tables S4 and S5, we find that the Williams features are somewhat better in risk stratification with the potential of enhancing personalized treatment. For example, associating new patients with LCA/Williams phenotype 1 patients that did not respond to IV fluids and will require vasopressors, as well as those that will require MV may inform early aggressive interventions and potentially improved clinical outcomes.[57–61] Moreover, the Williams criteria allowed for the use of race in risk stratification, with results indicating that non-Caucasian patients have a higher probability of being assigned to a high-severity phenotype compared to Caucasian patients. When the Seymour parameters were used, such separation was not observed.

On the other hand, the K-means/Seymour phenotypes (Supporting Information Table S7) showed some discriminating ability for the less severe cases which can be helpful in a screening process for the clinicians when deciding disposition. However, this model did not perform as well in risk stratification and the phenotypes obtained were not very distinct in terms of disease severity and specific organ dysfunction. Moreover, although minimizing number of subgroups is optimized with K-means compared to LCA, the variability from 3 to 4 components using K-means (Supporting Information Table S6) versus LCA suggested diminished reproducibility of K-means across varying feature sets. A drawback of LCA compared to K-means is that LCA is computationally demanding.[62] This can be a limitation on the number of cases and features used to derive clusters using LCA. Currently, the upper limit on how big the data can be for LCA modeling is unknown and depends on the available processing power.[17]

The lack of a higher degree of agreement in phenotypes between the 2 sets of LCA segmentations (Williams vs Seymour criteria) is mainly due to the difference in the focus of specific organ dysfunction. The Seymour criteria include inflammatory markers such as ESR and CRP but also more specific cardiac and liver biomarkers, because they are targeted to the adult population. On the other hand, the Williams criteria has more parameters for lung function, which is relevant in pediatric sepsis, which tends to present with more prominent respiratory symptoms. ESR and CRP are non-specific inflammatory markers that are not as important in pediatric sepsis and have largely been replaced by other better performing biomarkers such as procalcitonin. Another interesting finding is that when transaminases are added as parameters, it seems that phenotype 4 in Seymour, which represents the sickest patients, is more closely related to phenotype 4 of the Williams analysis, despite the fact that the latter includes patients that are less ill than phenotype 1. This discrepancy is caused by the varying baseline levels of hepatic dysfunction and can be explained by the fact that our pediatric population included many patients with history of liver transplantation that underlines the importance of potentially incorporating baseline values into the analysis.

4.1 | Provider-guided sepsis diagnosis and treatment align with the phenotypic analysis

Discrepancies between severe sepsis/shock diagnoses and treatments within phenotypes were observed, likely due to the ambiguous/evolving pediatric sepsis consensus diagnostic criteria and reflecting early sepsis patients who received sepsis treatment in efforts to avert organ dysfunction. It is therefore important to note that many of the septic patients with varying degrees of organ dysfunction might not technically meet the sepsis criteria determined by consensus but

are still at an increased risk for sequelae and were correctly clustered by our LCA model. We were also able to identify patients that were justifiably deemed to have severe sepsis by the clinicians because of their underlying complex medical or surgical conditions and the presence of a severe infection but without meeting the standard criteria of organ dysfunction.

4.2 | Moving toward personalized therapeutics

These findings suggest that clustering of highly heterogeneous clinical pediatric sepsis observations to find subgroupings of patients, analogous to diagnostic categories, may provide useful prognostic information. Furthermore, phenotype membership can be informative of the therapies that are likely to be given to similar patients, which suggests that this type of analysis, if prospectively validated, may help elucidate potential individual therapies. It is interesting to note that unlike bandemia that statistically differed phenotypes and was associated with severity,[63] traditional inflammatory biomarkers such as CRP and ESR were not statistically different among phenotypes. More effective markers such as procalcitonin are slowly leading to the replacement of CRP and ESR in pediatric sepsis diagnostics.[64]

Although more studies are needed in the pediatric population, the ability to phenotype patients early in the course of their disease and predict a more severe clinical course may present an opportunity for personalized therapeutics to alleviate disease sequelae. Given the ongoing concerns regarding the use of aggressive fluid management in some patients,[23,65] we believe the identification of these phenotypes may help triage pediatric sepsis patients that respond differently to aggressive fluids treatment,[13,66] predict individualized responses to certain medications or interventions (eg, normal saline vs Ringer's lactate fluids), and help identify those who will need vasopressors, potentially avoiding circulatory collapse and mortality.

The findings of this study that derived clinically useful phenotypic separation of septic children is also significant since complex pediatric patients that are appropriate targets for early management are identified despite not yet exhibiting life-threatening organ dysfunctions.

In summary, clustering techniques using routinely available EHR data can lead to clinically useful phenotyping identification in the pediatric sepsis-severe sepsis continuum. In this dataset of children with highly mixed forms of sepsis, as compared to K-means, LCA resulted in superior partitions of sepsis severity, treatments, and outcomes that were non-randomly distributed across phenotypes. These experiments suggest that LCA combined with predictive modeling may be useful in real-time analysis of EHR data collected in the ED setting toward identification of pediatric sepsis phenotypes to inform personalized care. Pilot studies are needed to validate the clinical use of EHR data clustering toward personalized therapies that improve outcomes.

ACKNOWLEDGMENT

Research reported in this publication was supported by a NIH SBIR award to Computer Technology Associates (CTA) by NIH National Institute of General Medical Sciences under award number 1R43GM122154-01. Additional analysis following the SBIR grant period of performance and development of this manuscript funded by CTA.

AUTHOR CONTRIBUTIONS

IK and TV performed as the co-principal investigators of the underlying NIH research study, conceived the current study, and authored major segments of the Introduction and Discussion sections of the manuscript. IK and TV, the guarantors of the article, take responsibility for the integrity of the work, from inception to published article. TW performed as the principal data scientist and LCA/machine learning analyst, developed models and authored major sections of Methods and Results sections of the manuscript. Clinicians RF and JC provided clinical support for the clinical interpretation of the results of the LCA analysis and edited the manuscript. Clinicians JG and SY provided clinical support in EHR data extraction and interpretation of results. All authors read and approved the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

ETHICS STATEMENT

The study has been approved by Advarra Institutional Board Review (MOD00413595). As records were deidentified, waiver of consent was granted.

ORCID

Ioannis Koutroulis MD, PhD https://orcid.org/0000-0002-8396-9022

REFERENCES

1. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. *Yearb Med Inform*. 2014;9(1):97.
2. Yan S, Kwan YH, Tan CS, Thumboo J, Low LL. A systematic review of the clinical application of data-driven population segmentation analysis. *BMC Med Res Methodol*. 2018;18(1):121.
3. Lanza ST, Rhoades BL. Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. *Prev Sci*. 2013;14(2):157-168.
4. Oberski D. Mixture Models: Latent Profile and Latent Class Analysis. In: Robertson J., Kaptein M. (eds) *Modern Statistical Methods for HCI*. Human Computer Interaction Series. Springer, Cham. 2016. https://doi.org/10.1007/978-3-319-26633-6_12
5. Boehmke B, Greenwell B. Hands-On Machine Learning with R (1st ed.). *Chapman and Hall/CRC*. (2019). https://doi.org/10.1201/9780367816377
6. Gårdlund B, Dmitrieva NO, Pieper CF, Finfer S, Marshall JC, Taylor Thompson B. Six subphenotypes in septic shock: latent class analysis of the PROWESS Shock study. *J Crit Care*. 2018;47:70-79.
7. Sinha P, Delucchi KL, Thompson BT, McAuley DF, Matthay MA, Calfee CS. Latent class analysis of ARDS subphenotypes: a secondary analysis of the statins for acutely injured lungs from sepsis (SAILS) study. *Intensive Care Med*. 2018;44(11):1859-1869.
8. Calfee CS, Delucchi K, Parsons PE, Thompson BT, Ware LB, Matthay MA. Subphenotypes in acute respiratory distress syndrome: latent class analysis of data from two randomised controlled trials. *Lancet Respir Med*. 2014;2(8):611-620.

9. Wilson JG, Calfee CS. ARDS subphenotypes: understanding a heterogeneous syndrome. *Crit Care*. 2020;24(1):102.
10. Zhang Z, Zhang G, Goyal H, Hong Y. Identification of subclasses of sepsis that showed different clinical outcomes and responses to amount of fluid resuscitation: a latent profile analysis. *Crit Care*. 2018;22(347):1-11.
11. Williams JB, Ghosh D, Wetzel RC. Applying machine learning to pediatric critical care data*. *Pediatr Crit Care Med*. 2018;19(7):599-608..
12. Seymour CW, Kennedy JN, Wang S, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA*. 2019;321:2003-2017.
13. Zhang Z, Zhang G, Goyal H, Mo L, Hong Y. Identification of subclasses of sepsis that showed different clinical outcomes and responses to amount of fluid resuscitation: a latent profile analysis. *Crit Care*. 2018;22(1):347. https://doi.org/10.1186/s13054-018-2279-3
14. Guilamet MCV, Bernauer M, Micek ST, Kollef MH. Cluster analysis to define distinct clinical phenotypes among septic patients with bloodstream infections. *Medicine (Baltimore)*. 2019;98(16):e15276..
15. Feuillet F, Bellanger L, Hardouin JB, Victorri-Vigneau C, Sébille V. On comparison of clustering methods for pharmacoepidemiological data. *J Biopharm Stat*. 2015;25(4):843-856.
16. Goodman LA. Latent class analysis: the empirical study of latent types, latent variables, and latent structures. In J. Hagenaars & A. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp. 3-55). Cambridge University Press; 2009. https://doi.org/10.1017/cbo9780511499531.002.
17. Sinha P, Calfee CS, Delucchi KL. Practitioner's guide to latent class analysis: methodological considerations and common pitfalls. *Crit Care Med*. 2021;49(1):e63-e79.
18. Paul R, Neuman MI, Monuteaux MC, Melendez E. Adherence to PALS sepsis guidelines and hospital length of stay. *Pediatrics*. 2012;130(2):e273-280..
19. Balamuth F, Weiss SL, Fitzgerald JC, et al. Protocolized treatment is associated with decreased organ dysfunction in pediatric severe sepsis. *Pediatr Crit Care Med*. 2016;17(9):817-822.
20. Lane RD, Funai T, Reeder R, Larsen GY. High reliability pediatric septic shock quality improvement initiative and decreasing mortality. *Pediatrics*. 2016; 138(4):e20154153.
21. Iskander KN, Osuchowski MF, Stearns-Kurosawa DJ, et al. Sepsis: multiple abnormalities, heterogeneous responses, and evolving understanding. *Physiol Rev*. 2013;93(3):1247-1288.
22. Singer M. Sepsis: personalization v protocolization?. *Crit Care*. 2019;23(1):127.
23. Cruz AT, Lane RD, Balamuth F, et al. Updates on pediatric sepsis. *J Am Coll Emerg Physicians Open*. 2020;1(5):981-993.
24. Ranjit S, Natraj R, Kandath SK, Kissoon N, Ramakrishnan B, Marik PE. Early norepinephrine decreases fluid and ventilatory requirements in pediatric vasodilatory septic shock. *Indian J Crit Care Med*. 2016;20(10):561-569.
25. Mesko B. The role of artificial intelligence in precision medicine. *Expert Rev Precis Med Drug Dev*. 2017;2(5):239-241.
26. Scott HF, Brilli RJ, Paul R, et al. Evaluating pediatric sepsis definitions designed for electronic health record extraction and multicenter quality improvement. *Crit Care Med*. 2020;48(10):e916-e926. https://doi.org/10.1097/CCM.0000000000004505. Published online 2020.
27. Asparouhov T. Variable-Specific Entropy Contribution. Published online 2014. https://www.semanticscholar.org/paper/Variable-Specific-Entropy-Contribution-Asparouhov/704f5d207e12386326501efd78c78b84267f4da5. Accessed January 13, 2022.
28. Magidson J, Vermunt J. Latent class models for clustering : a comparison with K-means. *Can J Mark Res*. 2002;20:37-44.
29. Pollack MM, Patel KM, Ruttimann UE. PRISM III: an updated pediatric risk of mortality score. *Crit Care Med*. 1996;24(5):743-752.
30. Study designs — Centre for Evidence-Based Medicine (CEBM) University of Oxford. https://www.cebm.ox.ac.uk/resources/ebm-tools/study-designs. Accessed August 15, 2021.
31. Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E. Epidemiology and costs of sepsis in the United States-An analysis based on timing of diagnosis and severity level. *Crit Care Med*. 2018;46(12):1889-1897.
32. Balamuth F, Weiss SL, Hall M, et al. Identifying pediatric severe sepsis and septic shock: accuracy of diagnosis codes. *J Pediatr*. 2015;167(6):1295-300.e4.
33. Sepanski RJ, Godambe SA, Mangum CD, Bovat CS, Zaritsky AL, Shah SH. Designing a pediatric severe sepsis screening tool. *Front Pediatr*. 2014;2(June):56.
34. Pollack MM, Patel KM, Ruttimann UE. PRISM III: an updated pediatric risk of mortality score. *Crit Care Med*. 1996;24(5):743-752.
35. Seymour CW, Kennedy JN, Wang S, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *JAMA - J Am Med Assoc*. 2019;321(20):2003-2017.
36. Goldstein B, Giroir B, Randolph A. International pediatric sepsis consensus conference: definitions for sepsis and organ dysfunction in pediatrics. *Pediatric Critical Care Medicine*. 2005;6(1):2-8.
37. Bishop CM. Mixture models and EM. In: Pattern recognition and machine learning. Springer, Cham. 2006. Published online.
38. Fraley C, Raftery AE. Model-based clustering, discriminant analysis, and density estimation. *J Am Stat Assoc*. 2002;97(458):611-631.
39. Scrucca L, Fop M, Murphy TB, Raftery AE. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J*. 2016;8(1):289-317.
40. Gower JC. A general coefficient of similarity and some of its properties. *Biometrics*. 1971;27(4):857-871.
41. Şenbabaoğlu Y, Michailidis G, Li J. Critical limitations of consensus clustering in class discovery. *Sci Rep*. 2014;27(4):6207.
42. Wilkerson MD, Hayes DN. ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. *Bioinformatics*. 2010;26(12):1572-1573.
43. Muthén LK, Muthén BO. How to use a Monte Carlo study to decide on sample size and determine power. *Struct Equ Model*. 2002;9(4):599-620.
44. Handl J, Knowles J, Kell DB. Computational cluster validation in post-genomic data analysis. *Bioinformatics*. 2005;21(15):3201-3212.
45. Brock G, Pihur V, Datta S, Datta S. ClValid: an R package for cluster validation. *J Stat Softw*. 2008;25(4):1-22.
46. Dunn JC. Well-separated clusters and optimal fuzzy partitions. *J Cybern*. 1974;4(1):95-104.
47. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987;20:53-65.
48. Model Based Clustering Essentials - Datanovia (online). https://www.datanovia.com/en/lessons/model-based-clustering-essentials/. Accessed May 15, 2021.
49. Faix JD. Biomarkers of sepsis. *Crit Rev Clin Lab Sci*. 2013;50(1):23-36.
50. Lubis MDS, Mawengkang H, Suwilo S. Performance analysis of entropy methods on K means in clustering process. *J Phys Conf Ser*. 2017;930:12028.
51. Gao HM, Ambroggio L, Shah SS, Ruddy RM, Florin TA. Predictive value of clinician "gestalt" in pediatric community-acquired pneumonia. *Pediatrics*. 2021;147(5):e2020041582.
52. Wong HR, Salisbury S, Xiao Q, et al. The pediatric sepsis biomarker risk model. *Crit Care*. 2012;16(5):R174.
53. Oikonomakou MZ, Gkentzi D, Gogos C, Akinosoglou K. Biomarkers in pediatric sepsis: a review of recent literature. *Biomark Med*. 2020;14(10):895-917..
54. Futoma J, Doshi-Velez F, Leo D, et al. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Heal*. 2020;2(9):e489-e492.
55. Singer M. Personalizing sepsis care. *Crit Care Clin*. 2018;34(1):153-160.

56. Weiss SL, Peters MJ, Alhazzani W, et al. Surviving sepsis campaign international guidelines for the management of septic shock and sepsis-associated organ dysfunction in children. *Intensive Care Med*. 2020;21(2):e52-e106. https://doi.org/10.1007/s00134-019-05878-6.
57. Hamzaoui O, Shi R. Early norepinephrine use in septic shock. *J Thorac Dis*. 2020;12(Suppl 1):S72-S77. https://doi.org/10.21037/jtd.2019.12.50.
58. Jozwiak M, Hamzaoui O, Monnet X, Teboul JL. Fluid resuscitation during early sepsis: a need for individualization. *Minerva Anestesiol*. 2018;84(8):987-992.
59. Ospina-Tascón GA, Hernandez G, Alvarez I, et al. Effects of very early start of norepinephrine in patients with septic shock: a propensity score-based analysis. *Crit Care*. 2020;24(1):52.
60. Russell JA, Gordon AC, Walley KR. Early may be better: early low-dose norepinephrine in septic shock. *Am J Respir Crit Care Med*. 2019;199(9):1049-1051.
61. Kohno S, Seki M, Takehara K, et al. Prediction of requirement for mechanical ventilation in community-acquired pneumonia with acute respiratory failure: a multicenter prospective study. *Respiration*. 2013;85(1):27-35.
62. Feuillet F, Bellanger L, Hardouin JB, Victorri-Vigneau C, Sébille V. On comparison of clustering methods for pharmacoepidemiological data. *J Biopharm Stat*. 2015;25(4):843-856.
63. Cavallazzi R, Bennin CL, Hirani A, Gilbert C, Marik PE. Is the band count useful in the diagnosis of infection? An accuracy study in critically ill patients. *J Intensive Care Med*. 2010;25(6):353-357.
64. Memar MY, Varshochi M, Shokouhi B, Asgharzadeh M, Kafil HS. Procalcitonin: the marker of pediatric bacterial infection. *Biomed Pharmacother*. 2017;96:936-943.
65. Byrne L, Van Haren F. Fluid resuscitation in human sepsis: time to rewrite history?. *Ann Intensive Care*. 2017; 7(1:4.
66. Watkins L. Interventions for pediatric sepsis and their impact on outcomes: a brief review. *Healthcare*. 2018;7(1):2.

AUTHOR BIOGRAPHY

Ioannis Koutroulis, MD, PhD, is the Assistant Director of Research in the Division of Emergency Medicine at Children's National Hospital in Washington, DC, USA.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Koutroulis I, Velez T, Wang T, et al. Pediatric sepsis phenotypes for enhanced therapeutics: An application of clustering to electronic health records. *JACEP Open*. 2022;3:e12660. https://doi.org/10.1002/emp2.12660