ABSTRACT

Introduction: Sepsis is a heterogeneous clinical syndrome. Identification of sepsis subphenotypes could lead to allowing more precise therapy. However, there is a lack of models to identify the subphenotypes in such patients. Thus, we aimed to identify possible subphenotypes and compare the clinical outcomes for subphenotypes in a large sepsis cohort.

Methods: This machine learning-based, cluster analysis was performed using the Medical Information Mart in Intensive Care (MIMIC)-IV database. We enrolled all adult (>18 years old) patients diagnosed with sepsis in the first 24 h after intensive care unit (ICU) admission. K-means cluster analysis was performed to identify the number of classes. Multivariable logistic regression models were used to estimate the association between sepsis subphenotypes and in-hospital mortality.

Results: A total of 8817 participants with sepsis were enrolled. The median age was 66.8 (IQR, 55.9–77.1) years, and 38.1% (3361/8817) were female. Two subphenotypes resulted in optimal separation including 11 routinely available clinical variables obtained during the first 24 h after ICU admission. Participants in subphenotype B showed higher levels of lactate, glucose and creatinine, white blood cell count, sodium and heart rate and lower body temperature, platelet count, systolic blood pressure, hemoglobin and PaO2/FiO2 ratio. In addition, the in-hospital mortality in patients with subphenotype B was significantly higher than that in subphenotype A (29.4% vs. 8.5%, P < 0.001). The difference was still significant after adjustment for potential covariates (adjusted OR 2.214; 95% CI 1.780–2.754, P < 0.001).

Conclusions: Two sepsis subphenotypes with different clinical outcomes could be rapidly identified using the K-means clustering analysis based on routinely available clinical data. This finding may help clinicians to identify the subphenotype rapidly at the bedside.
Graphical abstract:

**Infectious Diseases and Therapy**

Application of Machine Learning for Clinical Subphenotype Identification in Sepsis

**A MACHINE LEARNING-BASED, CLUSTER ANALYSIS**

**Question**

Can machine learning algorithms be applied to identify possible sepsis subphenotypes using routinely available clinical data?

**Database and population**

MIMIC-IV 2008-2019

Sepsis N = 8817

Age: 66.8 years

Female: 38.1%

**Methods**

Before Grouping Clusters

Cluster analysis

After Grouping Clusters

**Results**

Two sepsis subphenotypes with different clinical outcomes can be rapidly identified using the K-means clustering analysis based on routinely available clinical data. This finding may help clinicians to rapidly and easily identify subphenotype of sepsis at the bedside.

The graphical abstract represents the opinions of the authors. For a full list of declarations, including funding and author disclosure statements, please see the full text online. © The authors, CC-BY-NC 2022

**Keywords:** Sepsis; Machine learning; Subphenotype; Critically illness; Precision medicine
Key Summary Points

**Why carry out this study?**
Sepsis is a heterogeneous clinical syndrome characterized by a dysregulated host response to infection.

Identifying the sepsis subphenotypes could lead to a better understanding of the pathophysiology and discovery of new treatment targets. However, there is a lack of models to identify the subphenotypes in such patients.

The aim of this study was to propose a machine-learning method to identify the sepsis subphenotype, using only routinely available clinical data collected within the first 24 h of ICU admission.

**What was learned from the study?**

Machine learning-based algorithms for subphenotype identification in sepsis are possible.

Two sepsis subphenotypes could be identified rapidly based on routinely available clinical data.

Subphenotype B was independently associated with increased in-hospital mortality among patients with sepsis.

DIGITAL FEATURES

This article is published with digital features, including a graphical abstract, to facilitate understanding of the article. To view digital features for this article go to https://doi.org/10.6084/m9.figshare.20418600.

INTRODUCTION

Sepsis is a common and frequently fatal clinical condition characterized by a dysregulated host response to infection [1]. Data from the recently epidemiologic studies suggested that the worldwide incidence of sepsis is estimated to 48.9 million cases per year, associated with mortality rates of 20–25% [2, 3]. Although the Surviving Sepsis Campaign Guidelines for Management of Sepsis and Septic Shock have undergone five updates within the last 2 decades since first introduced in 2004, the mortality rate among patients with sepsis remains unacceptably high [4]. A major potential barrier to progress is the heterogeneity in sepsis.

Not all sepsis is the same. However, up to now, a one-size-fits-all approach is still being implemented for clinical practice, which ignores the heterogeneity across sepsis patients. Recently, several studies have accurately identified subphenotypes among sepsis cases; these subphenotypes have different demographics, laboratory values and clinical outcomes [5–8]. Notably, these methods for subphenotype identification are largely reliant on the measurement of specific protein biomarkers, such as vascular adhesion protein 1 (VAP1), matrix metalloproteinase 8 (MMP8) and proteinase 3 (PRTN3) [7]. However, these variables are not widely available as a routine clinical test, and the high prices for biomarker detection limit the rapid identification of the sepsis subphenotypes in clinical practice. Thus, there is a need to derive the sepsis subphenotypes by using routinely available clinical data in the early intensive care unit (ICU) admission stage.

Machine-learning classifier models that use clinical data could be performed to identify disease subphenotypes. Among them, K-means cluster analysis is a good clustering method that has already gained a wide range of acceptability in medicine [9, 10]. The K-means cluster-based methods have been extensively applied for subphenotype identification of pediatric acute respiratory distress syndrome (ARDS) [11] and sepsis-associated acute kidney injury [12]. However, to the best of our knowledge, there is a lack of models to identify the clinical subphenotypes in patients with sepsis.

Accordingly, the objective of this study was to propose a K-means cluster method to identify the sepsis subphenotype, using only routinely available clinical data collected within the first 24 h of ICU admission.
METHODS

Ethical Approval

The data for this study were obtained from the Medical Information Mart for Intensive Care (MIMIC-IV) database. The establishment of the MIMIC-IV database was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived because this project did not impact clinical care, and all protected health information was deidentified [13]. This study was performed according to the Declaration of Helsinki in 2013 [14] and reported in accordance with the Strengthening the Reporting of Observational studies in Epidemiology (STROBE) statement [15].

Data Source

As an update lately to the MIMIC-III, the current MIMIC-IV (version 1.0) is a large, freely available, open-access database comprising a variety of clinical-related data associated with 76,540 distinct admissions for patients who stayed in the ICU of the Beth Israel Deaconess Medical Center between 2008 and 2019 [13]. After completing a recognized course in protecting human research participants and signing a data use agreement, one author (Chang Hu) was approved to access the database (certification no. 47460147).

Study Population and Outcome

In this study, sepsis was defined as a confirmed or suspected infection combined with a Sequential Organ Failure Assessment (SOFA) score ≥ 2 in the first 24 h after ICU admission, in accordance with the Third International Consensus Definitions for Sepsis in 2016 [1]. The timeline of sepsis definition was present in Supplementary Fig. S1. We enrolled all adult (> 18 years) septic patients. Exclusion criteria were multiple ICU admissions (only the data of each patient’s first ICU admission were used in this study) and ICU length of stay (LOS) < 24 h. The primary outcome was in-hospital mortality; the secondary outcomes were ICU mortality, ICU length of stay (LOS) and hospital LOS.

Data Extraction and Preprocessing

In the MIMIC-IV database, we extracted a set of clinical variables. As we have previously described [16], these were routinely accessed parameters, including demographic variables (e.g., age, gender, ethnicity, body weight, height, admission time period and admission type), medical history (e.g., hypertension, diabetes, congestive heart failure, cerebrovascular disease, chronic pulmonary disease, liver disease, renal disease, tumor and acquired immune deficiency syndrome), vital signs (e.g., heart rate, systolic blood pressure, diastolic blood pressure, mean arterial pressure, respiratory rate, body temperature and SpO2), laboratory findings (e.g., blood glucose, lactate, pH, pCO2, pO2, base excess, white blood cell, anion gap, bicarbonate, blood urea nitrogen, serum calcium, serum chloride, serum creatinine, serum sodium, serum potassium, serum fibrinogen, international normalized ratio, prothrombin time, partial thromboplastin time, alanine aminotransferase, alkaline phosphatase, aspartate aminotransferase, total bilirubin, amylase,
### Table 1  Demographics at baseline

| Characteristics          | Total (N = 8817) | Subphenotype A (n = 7094) | Subphenotype B (n = 1723) | P value |
|--------------------------|------------------|---------------------------|---------------------------|---------|
| **Demographics**         |                  |                           |                           |         |
| Age, years               | 66.8 (55.9–77.1) | 66.7 (55.7–77.1)          | 67.3 (56.6–77.3)          | 0.463   |
| Gender                   |                  |                           |                           | 0.691   |
| Male, n (%)              | 5456 (61.9)      | 4397 (62.0)               | 1059 (61.5)               |         |
| Female, n (%)            | 3361 (38.1)      | 2697 (38.0)               | 664 (38.5)                |         |
| Race and ethnicity       |                  |                           |                           | < 0.001 |
| White                    | 5887 (66.8)      | 4848 (68.3)               | 1039 (60.3)               |         |
| Black/African American   | 467 (5.3)        | 362 (5.1)                 | 105 (6.1)                 |         |
| Hispanic/Latinx          | 241 (2.7)        | 181 (2.6)                 | 60 (3.5)                  |         |
| Asian                    | 205 (2.3)        | 164 (2.3)                 | 41 (2.4)                  |         |
| Other                    | 462 (5.2)        | 375 (5.3)                 | 87 (5.0)                  |         |
| Unknown                  | 1555 (17.6)      | 1164 (16.4)               | 391 (22.7)                |         |
| Weight, kg               | 81 (69–95)       | 80 (68–95)                | 83 (70–97)                | < 0.001 |
| Height, cm               | 170 (163–178)    | 170 (163–178)             | 170 (163–178)             | 0.482   |
| Admission period, n (%)  |                  |                           |                           | < 0.001 |
| 2008–2010                | 3075 (34.9)      | 2553 (36.0)               | 522 (30.3)                |         |
| 2011–2013                | 2507 (28.4)      | 2065 (29.1)               | 442 (25.7)                |         |
| 2014–2016                | 2297 (26.1)      | 1836 (25.9)               | 461 (26.8)                |         |
| 2017–2019                | 938 (10.6)       | 640 (9.0)                 | 298 (17.3)                |         |
| Admission type, n (%)    |                  |                           |                           | < 0.001 |
| CCU                      | 3838 (43.5)      | 3085 (43.5)               | 753 (43.7)                |         |
| MICU                     | 1462 (16.6)      | 1067 (15.0)               | 395 (22.9)                |         |
| SICU                     | 1183 (13.4)      | 1029 (14.5)               | 154 (8.9)                 |         |
| TSICU                    | 1123 (12.7)      | 929 (13.1)                | 194 (11.3)                |         |
| MICU/SICU                | 1041 (11.8)      | 823 (11.6)                | 218 (12.7)                |         |
| Others                   | 170 (1.9)        | 161 (2.3)                 | 9 (0.5)                   |         |
| Comorbidities            |                  |                           |                           |         |
| Charlson comorbidity index | 5 (4–7)         | 5 (3–7)                   | 6 (4–8)                   | < 0.001 |
| Hypertension, n (%)       | 4256 (48.3)      | 3512 (49.5)               | 744 (43.2)                | < 0.001 |
| Diabetes, n (%)           | 2528 (28.7)      | 1882 (26.5)               | 646 (37.5)                | < 0.001 |
| Congestive heart failure, n (%) | 2365 (26.8) | 1773 (25.0) | 592 (34.4) | < 0.001 |
| Cerebrovascular disease, n (%) | 1234 (14.0) | 1010 (14.2) | 224 (13.0) | 0.184 |
creative phosphokinase, creatine kinase MB, lactate dehydrogenase, PaO2/FiO2 ratio, hematocrit, hemoglobin, platelets and albumin), medical treatments (e.g., mechanical ventilation, the time to first dose of antibiotic agents and vasopressor use), urine output and Glasgow Coma Scale score. For each variable, we extracted the most abnormal value recorded within the first 24 h of ICU admission. Missingness is considered to be missing at random, and variables with > 30% missing values were excluded from the analysis. Furthermore, we employed the multiple imputation by chained equations (MICE) method to handle the remaining missing data (MICE package in R).

### Variable Selection

In feature selection part, we followed the procedure of Soussi et al. [6] and Zhang et al. [8]. In brief, we first removed the variables with a missing fraction > 30%. Then, we captured the preselected variables based on the prior published literature and their potential association with sepsis onset and progress. The final selection of variables included in the K-means clustering algorithm (11 variables) was made by consensus among two critical care medicine experts: YL and ZP (Supplementary Table S1). The PaO2/FiO2 ratio was used for respiratory function; serum creatinine was used for renal function; platelet and hemoglobin were used for hematologic system; the lactate, heart rate, systolic blood pressure and body temperature were used for circulatory system; white blood cell count was used for inflammatory condition; sodium was used for electrolyte parameters; glucose was used for metabolism function.

### Subphenotype Classification

In the present study, we employed a K-means clustering algorithm to determine clusters. We first assessed the correlation between candidate variables. Then, all continuous variables were transformed into z-score (mean: 0, standard deviation: -1 to 1) in the algorithm. Two to eight clusters were compared in this unsupervised approach. We determined the optimal number of clusters based on the analysis of total within sum of squares (WSS), the Silhouette score (ranged from -1 to 1, a value closer to 1 being better), Davies-Bouldin score (ranged from 0 upward, a value closer to 0 being better) and Calinski-Harabasz score (ranged from 0 upward, higher value being better). Finally, we used a principal component analysis (PCA) to visualize the clustering results.

### Statistical Analysis

Categorical variables were expressed as number and percentage and tested for baseline differences.
| Clinical characteristics                      | Total \((N = 8817)\) | Subphenotype A \((n = 7094)\) | Subphenotype B \((n = 1723)\) | \(P\) value |
|---------------------------------------------|----------------------|-------------------------------|-------------------------------|-------------|
| **Vital signs on day 1**                    |                      |                               |                               |             |
| Heart rate, bpm                             | 84 (75–95)           | 83 (74–93)                    | 89 (80–102)                   | < 0.001     |
| Systolic blood pressure, mmHg               | 114 (106–124)        | 116 (107–126)                | 109 (103–118)                 | < 0.001     |
| Diastolic blood pressure, mmHg              | 60 (55–67)           | 60 (55–67)                   | 59 (53–65)                    | < 0.001     |
| Mean arterial pressure, mmHg                | 76 (71–82)           | 76 (71–83)                   | 74 (69–80)                    | < 0.001     |
| Respiratory rate                            | 18 (16–21)           | 18 (16–21)                   | 20 (17–23)                    | < 0.001     |
| Body temperature, °C                        | 36.8 (36.6–37.2)     | 36.9 (36.6–37.2)             | 36.7 (36.4–37.1)              | < 0.001     |
| \(\text{SpO}_2\), %                        | 93 (91–95)           | 93 (91–95)                   | 92 (89–95)                    | < 0.001     |
| **Laboratory findings on day 1**            |                      |                               |                               |             |
| Blood glucose, mg/dl                        | 171 (141–211)        | 163 (135–192)                | 232 (188–317)                 | < 0.001     |
| Lactate, mmol/l                             | 2.3 (1.5–3.3)        | 2.0 (1.4–2.8)                | 4.8 (3.1–7.4)                 | < 0.001     |
| pH                                          | 7.43 (7.39–7.47)     | 7.43 (7.39–7.47)             | 7.42 (7.37–7.46)              | < 0.001     |
| \(\text{pO}_2\), mmHg                       | 85 (59–119)          | 91 (66–126)                  | 63 (41–85)                    | < 0.001     |
| \(\text{pCO}_2\), mmHg                      | 46 (41–53)           | 46 (41–51)                   | 50 (43–58)                    | < 0.001     |
| \(\text{PaO}_2/\text{FiO}_2\) ratio         | 185 (111–272)        | 202 (130–285)                | 113 (74–183)                  | < 0.001     |
| Base excess                                  | 1.0 (0.0–3.0)        | 1.0 (0.0–3.0)                | 0 (−3 to 2)                   | < 0.001     |
| Hematocrit, %                               | 28.9 (25.0–33.6)     | 29.5 (25.6–34.0)             | 26.1 (22.6–31.1)              | < 0.001     |
| Hemoglobin, g/dl                            | 9.7 (8.4–11.3)       | 9.9 (8.6–11.4)               | 8.8 (7.6–10.4)                | < 0.001     |
| Platelets, \(\times 10^3/\text{ul}\)       | 156 (111–217)        | 164 (119–223)                | 124 (80–181)                  | < 0.001     |
| White blood cell, \(\times 10^3/\text{ul}\) | 14.0 (10.4–18.6)     | 13.3 (10.1–17.4)             | 17.9 (13.1–24.4)              | < 0.001     |
| Anion gap                                    | 15 (13–18)           | 14 (12–17)                   | 19 (15–23)                    | < 0.001     |
| Bicarbonate, mmol/l                         | 24 (22–27)           | 25 (23–27)                   | 23 (20–26)                    | < 0.001     |
| Blood urea nitrogen, mg/dl                  | 20 (15–31)           | 19 (14–27)                   | 29 (19–49)                    | < 0.001     |
| Serum calcium, mmol/l                       | 8.4 (8.0–8.9)        | 8.4 (8.0–8.9)                | 8.5 (8.0–9.1)                 | < 0.001     |
| Serum chloride, mmol/l                      | 108 (104–111)        | 108 (104–111)                | 108 (104–112)                 | < 0.001     |
| Serum creatinine, mg/dl                     | 1.0 (0.8–1.5)        | 1.0 (0.8–1.3)                | 1.6 (1.1–2.8)                 | < 0.001     |
| Serum sodium, mmol/l                        | 140 (138–142)        | 140 (138–142)                | 141 (138–144)                 | < 0.001     |
| Serum potassium, mmol/l                     | 4.4 (4.1–4.9)        | 4.4 (4.1–4.8)                | 4.8 (4.4–5.4)                 | < 0.001     |
| International normalized ratio              | 1.3 (1.2–1.6)        | 1.3 (1.1–1.5)                | 1.5 (1.3–2.0)                 | < 0.001     |
comparability with the chi-square test or Fisher’s exact test, as appropriate. Continuous variables were expressed as mean and standard deviation or median and interquartile range (IQR) and compared between the two groups using the Student’s t-test or the Mann-Whitney test, as appropriate.

Multivariable logistic regression models were used to estimate the association between sepsis subphenotypes and in-hospital mortality. Unadjusted and adjusted odd ratios (ORs) and 95% confidence intervals (CIs) were calculated using multiple logistic regression models. In the unadjusted model, we tested the direct effect of sepsis subphenotypes on in-hospital mortality. Model 1 was adjusted for demographic variables (age, gender, body weight, height, racial, admission time period and admission type). Model 2 was adjusted for the covariates included in model 1 as well as the comorbidities (Charlson comorbidity index). Model 3 was adjusted for the covariates included in model 2 as well as the neurologic function (GCS). Model 4 was adjusted for the variables in model 3 along with the severity of illness score (SOFA). Model 5 incorporated model 4 along with

| Clinical characteristics | Total (N = 8817) | Subphenotype A (n = 7094) | Subphenotype B (n = 1723) | P value |
|--------------------------|-----------------|---------------------------|---------------------------|---------|
| Prothrombin time, s      | 14.7 (13.0–17.0) | 14.4 (12.8–16.3)          | 16.7 (14.3–21.8)          | < 0.001 |
| Partial thromboplastin time, s | 32.9 (28.4–42.9) | 31.9 (28.1–39.2)          | 41.1 (31.6–68.4)          | < 0.001 |
| Medical treatments on day 1 |                |                           |                           |         |
| Antibiotics, n (%)       | 6307 (71.5)     | 5028 (70.9)               | 1279 (74.2)               | 0.006   |
| Mechanical ventilation, n (%) | 5242 (59.5)     | 4042 (57.0)               | 1200 (69.6)               | < 0.001 |
| Vasopressors, n (%)      | 3510 (39.8)     | 2477 (34.9)               | 1033 (60.0)               | < 0.001 |
| Urine output on day 1, ml | 1690 (1085–2470) | 1770 (1206–2517)          | 1242 (587–2145)           |         |
| Severity of illness scores |                |                           |                           |         |
| GCS                      | 14 (9–15)       | 14 (10–15)                | 13 (6–14)                 | < 0.001 |
| SOFA                     | 5 (4–8)         | 5 (3–7)                   | 10 (7–13)                 | < 0.001 |
| OASIS                    | 35 (29–41)      | 33 (28–39)                | 41 (34–48)                | < 0.001 |
| LODS                     | 5 (3–8)         | 4 (3–7)                   | 8 (6–11)                  | < 0.001 |
| SAPS-II                  | 37 (29–47)      | 35 (27–43)                | 48 (38–60)                | < 0.001 |
| Primary outcome          |                |                           |                           |         |
| In-hospital mortality    | 1107 (12.6)     | 601 (8.5)                 | 506 (29.4)                | < 0.001 |
| Secondary outcomes       |                |                           |                           |         |
| ICU mortality            | 861 (9.8)       | 424 (6.0)                 | 437 (25.4)                | < 0.001 |
| ICU LOS                  | 3.0 (1.7–6.0)   | 2.8 (1.5–5.3)             | 4.6 (2.4–9.7)             | < 0.001 |
| Hospital LOS             | 8.2 (5.1–14.2)  | 7.9 (5.1–13.1)            | 10.6 (5.5–18.9)           | < 0.001 |

Data were reported as no. (%) or median (IQR)

GCS Glasgow Coma Scale, SOFA Sequential Organ Failure Assessment, OASIS Oxford Acute Severity of Illness Score, SAPS-II Simplified Acute Physiology Score II, LOS length of stay
medical treatments (antibiotic therapy on day 1, mechanical ventilation on day 1 and vasoressor use on day 1).

A two-sided $P$ value $< 0.05$ was considered statistically significant. All analyses were carried out using SPSS statistical software version 24.0 (IBM), R statistical software version 3.6.1 (R Foundation) and Python software version 3.6 (Fig. 1).

RESULTS

Characteristics of Cohorts

In the current analysis, 12,292 patients were screened with sepsis in the MIMIC-IV database. After all exclusions (2657 patients with multiple ICU admission and 818 patients whose LOS was $< 24$ h in the ICU), 8817 participants were finally enrolled in this study (Supplementary Fig. S2). Of the 8817 cases, the median age was 66.8 years (IQR 55.9–77.1 years), and 38.1% (3361/8817) were female. The most frequently documented race and ethnicity category was White (5887/8817, 66.8%), followed by Black/African American (467/8817, 5.3%), Hispanic/Latinx (241/8817, 2.7%) and Asian (205/8817, 2.3%). The top three most frequently reported comorbidities included hypertension (4256/8817, 48.3%), diabetes (2528/8817, 28.7%) and congestive heart failure (2365/8817, 26.8%) (Table 1). The demographics at baseline are provided in Table 1.

Among 8817 patients, 71.5% (6307/8817) were administered the first antibiotic dose in the first 24 h after ICU admission, 59.5% (5242/8817) received mechanical ventilation on day 1, and 39.8% (3510/8817) were given first vasoressor dose in the first 24 h after ICU admission.

![Correlation matrix of the variables measured in this cohort. Coefficients are derived using the Spearman’s rank correlation coefficient. *WBC* white blood cell, *SBP* systolic blood pressure](image)

**Fig. 2**
admission. The median SOFA score was 5 (IQR: 5–8) across the whole dataset, a result indicating more severe illness. The overall all-cause in-hospital mortality rate was 12.6% (1107/8817) (Table 2).

Derivation of Sepsis Subphenotypes

Supplementary Table 1 presents the 11 clinical variables covering the functions of several organs. Significant correlations are also noted between the candidate variables (Fig. 2). The highest correlations were observed between glucose and lactate ($r = 0.3$) and heart rate and body temperature ($r = 0.3$). Supplementary Table S2 displays a summary of $K$-means clustering ranging from 2 to 7 for this cohort, respectively. Clustering with $K = 2$ was found to have a higher Silhouette score of 0.24 and a higher Calinski-Harabasz score of 879 compared with other classes. The within-cluster variance graph is presented in Supplementary Fig. S3. Therefore, a two-class model provided the optimal fit in this study. For simplicity, we henceforth refer to the two classes as subphenotypes A ($N = 7094$) and B ($N = 1723$), respectively. For easier exploration and visualization of two subphenotypes, we also created two-dimensional images using principal component analysis (PCA) to mark the differences between before and after clustering (Fig. 3).

Characteristics of Each Subphenotype

Figure 4 shows the selected variables for the two subphenotypes. Compared with subphenotype A, subphenotype B was defined by considerably higher levels of lactate, glucose, creatinine, white blood cell count and sodium; higher heart rate; and lower body temperature, platelet, systolic blood pressure, hemoglobin and PaO$_2$/FiO$_2$ ratio. Specifically, the details for comparison of these variables between the two subphenotypes are presented in Fig. 5.

Baseline characteristics of study participants according to subphenotype are also shown in Tables 1 and 2. There were no significant differences in age, gender and height between subphenotype A and subphenotype B. Subjects in subphenotype A vs subphenotype B were more likely to be White (68.3% vs. 60.3%). In addition, participants in subphenotype A vs
subphenotype B had lower Charlson Comorbidity Index (5 [3–7] vs. 6 [4–8], P < 0.001) and lower need for antibiotic therapy (70.9% vs. 74.2%, P < 0.001), mechanical ventilation (57.0% vs. 69.6%, P < 0.001) and vasopressor use (34.9% vs. 60.0%, P < 0.001).

Clinical Outcomes of Each Subphenotype

Results for all outcomes are shown in Table 2. The in-hospital and ICU mortality in subphenotypes B were significantly higher than those in subphenotype A (29.4% vs. 8.5%, P < 0.001; 25.4% vs. 6.0%, P < 0.001; respectively). Furthermore, the lengths of hospital stay and ICU stay were significantly longer in patients with subphenotype B compared with those in subphenotype A (10.6 [5.5–18.9] vs. 7.9 [5.1–13.1], P < 0.001; 4.6 [2.4–9.7] vs. 2.8 [1.5–5.3], P < 0.001; respectively).

Results of the univariable and multivariable logistic regression analysis for the primary outcome are presented in Table 3. In a univariable analysis, subjects in subphenotype B were associated with increased in-hospital mortality (OR 4.492; 95% CI 3.932–5.132; P < 0.001). After adjusting for multiple potential confounders using several multivariate logistic regression models, we found that subphenotype B was independently associated with increased risk of in-hospital mortality compared with subphenotype A (adjusted OR 2.214; 95% CI 1.780–2.754, P < 0.001).

DISCUSSION

In this study, we demonstrated that K-means cluster analysis, using only routinely available clinical data as factors, could accurately identify the sepsis subphenotypes. We captured 11 representative and easily accessible variables related to different organ systems and observed important differences between the two identified subphenotypes. Additionally, patients in subphenotype B had significant higher mortality even after adjusting for potential confounders compared with those in subphenotype A. Taken together, this finding might be a valuable tool for prognostication stratification of sepsis in the clinical practice.

Data-driven cluster analysis is widely used for diseases classification [17]. Among several cluster analysis methods, the K-means algorithm is the most popular machine-learning clustering algorithm [18]. To date, several
Fig. 5 Comparison of the selected variables between the two subphenotypes in sepsis
researchers have successfully applied K-means cluster analysis to identify asthma phenotypes [17], Parkinson’s disease subtypes [19] and complex regional pain syndrome phenotypes [20]. Thus, the K-means cluster analysis used in this study seemed to be an appropriate choice since it maximizes separation between clusters and offering the greatest scope for identifying distinct groups within the patients [16].

Identification of distinct subphenotypes in sepsis is a key component of personalized medicine, which may help in better risk stratification and treatment decisions. Notably, how to translate research into clinical practice remains one of the biggest challenges in subphenotypes identification [10]. For example, HBP (heparin-binding protein), Ela (neutrophil elastase 2), PRTN3 and MMP8 have been reported to be the key factors in phenotype identification for septic acute kidney injury [7]. However, these variables were not widely available as a routine clinical test, which failed to translate into clinically useful applications. In the current study, we used 11 routinely available clinical features for deriving the clinical phenotypes of sepsis in the first 24 h after ICU admission. The value of these parameters reflected the state of different target organs. These findings allowed the realization of personalized physiologic medicine to be practiced at the bedside for critically ill patients with sepsis.

Understanding of the underlying pathophysiologic mechanisms behind the subphenotype identification could help identify the subphenotype in patients with sepsis. Of the two subphenotypes identified in sepsis, subphenotype B was associated with higher levels of lactate, glucose and creatinine and lower levels of hemoglobin and PaO₂/FiO₂ ratio. Lactate is a marker of abnormal microcirculation [21], reflective of tissue hypoperfusion and cellular hypoxia [22], and has been reported to be a sensitive indicator to predict prognosis in sepsis [23]. The glycometabolism disorder is common in critically ill patients, especially those with sepsis [24]. A mild elevation of glucose is acceptable because it allows the host to survive during severe stress. However, an excessively high level of glucose may cause immunosuppression and oxidative stress, which were associated with worse outcomes [25, 26]. Serum

| Regression Models | Multivariable |
|-------------------|---------------|
|                   | OR (95% CI)   | P value |
| Unadjusted model  | 4.492 (3.932–5.132) | < 0.001 |
| Model 1           | 4.556 (3.825–5.427) | < 0.001 |
| Model 2           | 4.208 (3.525–5.024) | < 0.001 |
| Model 3           | 3.799 (3.166–4.558) | < 0.001 |
| Model 4           | 2.163 (1.741–2.688) | < 0.001 |
| Model 5           | 2.214 (1.780–2.754) | < 0.001 |

In the unadjusted model, we tested the direct effect of sepsis subphenotypes (subphenotype A vs. subphenotype B) on in-hospital mortality. Model 1 was adjusted for demographic variables (age, gender, body weight, height, racial, admission time period and admission type). Model 2 was adjusted for the covariates included in model 1 as well as the comorbidities (Charlson comorbidity index). Model 3 was adjusted for the covariates included in model 2 as well as the neurologic function (GCS). Model 4 was adjusted for the variables in model 3 along with the severity of illness score (SOFA). Model 5 incorporated model 4 along with medical treatments (antibiotic therapy on day 1, mechanical ventilation on day 1 and vasopressor use on day 1).
Creatinine is the most widely used measure of renal function in clinical practice. The significantly elevated creatinine in subphenotype B revealed the subjects in this group had a higher proportion of renal dysfunction compared with patients in subphenotype B [27]. This result is in accord with a recent study in which the serum creatinine was a predictor of mortality in sepsis [6]. Additionally, individuals in subphenotype B were found to have a lower PaO2/FiO2 ratio compared with those in subphenotype A. Previous study also demonstrated that the PaO2/FiO2 was an important indicator in respiratory function and widely used to differentiate groups of patients at high risk for adverse clinical outcomes [28]. Taken together, it is not difficult to understand why patients in subphenotype B had worse clinical outcomes compared to those in subphenotype A.

In the current study, we used a classification approach based on routinely available clinical data to yield insights into sepsis subphenotypes. Embedding this classification model into the electronic health record (EHR) system would allow for rapid bedside screening. Moreover, automating the capture and processing of an abundant data stream in the ICU would contribute to evaluating prognostic or therapeutic differences among septic patients. However, this advanced method for sepsis classification would require additional external validation before clinical implementation.

This study has several limitations. First, because the etiology of sepsis is complex, it might be not enough to discover all the subphenotypes using known rules and some important variables. Additionally, the MIMIC-IV database covers the period from 2008 to 2019 without the exact year of patient admission, and changes in the management of sepsis may have occurred in the interim, which increased the bias. We have attempted to partly mitigate this effect by applying a unified definition of sepsis (Sepsis-3) and divided patients into four groups in terms of admission time period (2008–2010, 2011–2013, 2014–2016 and 2017–2019) for enrollment in the model. The results held true even after adjusting for admission time period. Second, this study only focused on baseline data with the most abnormal value in the first 24 h of ICU admission, which limited dynamic classification for sepsis subphenotype. On the other hand, the natural progression of sepsis over time might lead to changing of subphenotype. Nonetheless, our static classification model could provide potential suggestions for care providers and prevent alert fatigue, which was blamed for high override rates in dynamic systems. Third, the inclusion of potential indicators may provide additional insights, such as type of infection, infection site, microbiology data, cigarette smoking data as well as drinking data. However, these were not available in this clinical database. Fourth, the lack of lactate and PaO2/FiO2 values was 22.18% and 27.39%, respectively. Although we used a multiple imputation approach to handle these missing data in the clustering model, this may still produce biased estimates of the relative risk. Fifth, sepsis was defined as a confirmed or suspected infection combined with a SOFA score \( \geq 2 \). However, we could not obtain the exact values except ICD-9 (or ICD-10) codes for chronic conditions in the electronic database. Therefore, we could not calculate the baseline SOFA score, which may have overestimated sepsis, and the results should be interpreted with caution. Sixth, the variables of clinical subphenotypes in this study were derived from a single-center retrospective database in the USA. Thus, it remains unknown whether these subphenotypes exist outside studies in a more diverse population of critically ill patients with sepsis. Additional studies are needed to further verify and validate the two distinct subphenotypes of sepsis.

**CONCLUSION**

Two sepsis subphenotypes with different clinical outcomes can be rapidly identified using the K-means clustering analysis based on routinely available clinical data. This finding may help clinicians to rapidly and easily identify the subphenotype of sepsis at the bedside.
ACKNOWLEDGEMENTS

We thank the participants of the study and the MIMIC-IV program for access to the database.

Funding. This work and the journal’s rapid service fee were funded by the National Natural Science Foundation of China (grants 81772046 and 81971816 to Dr. Peng) and the Subject Cultivation Project of Zhongnan Hospital of Wuhan University (Zhiyong Peng, no. ZNXKPY2021001).

Authorship. All named authors meet the International Committee of Medical Journal Editors (ICMJE) criteria for authorship for this article, take responsibility for the integrity of the work as a whole, and have given their approval for this version to be published.

Author Contributions. CH and ZP designed this study; CH and YL were responsible for the data collection; YL and FW were responsible for data analysis; CH and ZP conducted the manuscript writing; ZP critically revised the manuscript. All authors read and approved the final manuscript.

Disclosures. Chang Hu, Yiming Li, Fengyun Wang and Zhiyong Peng have nothing to disclose.

Compliance with Ethics Guidelines. The data for this study were obtained from the Medical Information Mart for Intensive Care (MIMIC-IV) database. The establishment of the MIMIC-IV database was approved by the Institutional Review Boards of Beth Israel Deaconess Medical Center (Boston, MA) and the Massachusetts Institute of Technology (Cambridge, MA). Requirement for individual patient consent was waived because this project did not impact clinical care and all protected health information was deidentified. This study was performed according to the Declaration of Helsinki in 2013 and reported in accordance with the Strengthening the Reporting of Observational studies in Epidemiology (STROBE) statement.

Data Availability. The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Open Access. This article is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License, which permits any non-commercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc/4.0/.

REFERENCES

1. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA. 2016;315(8):801–10.
2. Rudd KE, Johnson SC, Agesa KM, et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. Lancet. 2020;395(10219):200–11.
3. Rhee C, Dantes R, Epstein L, et al. Incidence and trends of sepsis in US hospitals using clinical vs claims data, 2009–2014. JAMA. 2017;318(13):1241–9.
4. Evans L, Rhodes A, Alhazzani W, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. Intensive Care Med. 2021;47(11):1181–247.
5. Seymour CW, Kennedy JN, Wang S, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. JAMA. 2019;321(20):2003–17.
6. Soussi S, Sharma D, Juni P, et al. Identifying clinical subtypes in sepsis-survivors with different one-year outcomes: a secondary latent class analysis of the FROG-ICU cohort. Crit Care. 2022;26(1):114.

7. Wiersema R, Jukarainen S, Vaara ST, et al. Two subphenotypes of septic acute kidney injury are associated with different 90-day mortality and renal recovery. Crit Care. 2020;24(1):150.

8. 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.

9. Sanchez-Pinto LN, Luo Y, Churpek MM. Big data and data science in critical care. Chest. 2018;154(5):1239–48.

10. Reddy K, Sinha P, O’Kane CM, Gordon AC, Calfee CS, McAuley DF. Subphenotypes in critical care: translation into clinical practice. Lancet Respir Med. 2020;8(6):631–43.

11. Yehya N, Varisco BM, Thomas NJ, Wong HR, Christie JD, Feng R. Peripheral blood transcriptomic sub-phenotypes of pediatric acute respiratory distress syndrome. Crit Care. 2020;24(1):681.

12. Chaudhary K, Vaid A, Duffy A, et al. Utilization of deep learning for subphenotype identification in sepsis-associated acute kidney injury. Clin J Am Soc Nephrol. 2020;15(11):1557–65.

13. Johnson AE, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Scientific data. 2016;3: 160035.

14. World Medical A. World Medical Association declaration of helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310(20):2191–4.

15. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Lancet. 2007;370(9596):1453–7.

16. Hu C, Li L, Huang W, et al. Interpretable machine learning for early prediction of prognosis in sepsis: a discovery and validation study. Infect Dis Ther. 2022;11(3):1117–32.

17. Haldar P, Pavord ID, Shaw DE, et al. Cluster analysis and clinical asthma phenotypes. Am J Respir Crit Care Med. 2008;178(3):218–24.

18. Selim SZ, Ismail MA. K-means-type algorithms: a generalized convergence theorem and characterization of local optimality. IEEE Trans Pattern Anal Mach Intell. 1984;6(1):81–7.

19. Pourzinal D, Yang JHJ, Byrne GJ, et al. Identifying subtypes of mild cognitive impairment in Parkinson’s disease using cluster analysis. J Neurol. 2020;267(11):3213–22.

20. Dimova V, Hermberger MS, Escolano-Lozano F, et al. Clinical phenotypes and classification algorithm for complex regional pain syndrome. Neurology. 2020;94(4):e357–67.

21. Ryoo SM, Lee J, Lee YS, et al. Lactate level versus lactate clearance for predicting mortality in patients with septic shock defined by sepsis-3. Crit Care Med. 2018;46(6):e489–95.

22. Zhai X, Yang Z, Zheng G, et al. Lactate as a potential biomarker of sepsis in a rat cecal ligation and puncture model. Mediators Inflamm. 2018;2018:8352727.

23. Scott HF, Brou L, Deakyne SJ, Kempe A, Fairclough DL, Bajaj L. Association between early lactate levels and 30-day mortality in clinically suspected sepsis in children. JAMA Pediatr. 2017;171(3):249–55.

24. Van den Berghe G, Wilmer A, Hermans G, et al. Intensive insulin therapy in the medical ICU. N Engl J Med. 2006;354(5):449–61.

25. Wang W, Chen W, Liu Y, et al. Blood glucose levels and mortality in patients with sepsis: dose-response analysis of observational studies. J Intensive Care Med. 2021;36(2):182–90.

26. Lu Z, Tao G, Sun X, et al. Association of blood glucose level and glycemic variability with mortality in sepsis patients during ICU hospitalization. Front Public Health. 2022;10: 857368.

27. De Rosa S, Samoni S, Ronco C. Creatinine-based definitions: from baseline creatinine to serum creatinine adjustment in intensive care. Crit Care. 2016;20:69.

28. Villar J, Perez-Mendez L, Blanco J, et al. A universal definition of ARDS: the PaO2/FiO2 ratio under a standard ventilatory setting—a prospective, multicenter validation study. Intensive Care Med. 2013;39(4):583–92.