

Clinically Contextualized State Spaces Provide Continuous Representations of Sepsis

Seibi Kobara, Morgan Kay, Tilendra Choudhary, Vijay Krishnamoorthy, Mihai V. Podgoreanu, Yoshiro Hayashi, Suresh M. Agarwal, and Rishikesan Kamaleswaran

Supplementary material

Supplementary Methods

Contextual dictionary and representation construction in Contextual Endotype and State Representation Learning (CERL)

Let $X \in \mathbb{R}^{n \times p}$ denote the input feature matrix for n observations and p biological or clinical features, and let $Y = \{Y_1, \dots, Y_K\}$ denote K prespecified clinical trajectory variables. For each trajectory k , we define a trajectory-specific contextual dictionary

$$\mathcal{D}_k = \{(\mathbf{x}_j, y_{jk})\}_{j=1}^n, \quad k = 1, \dots, K$$

where $\mathbf{x}_j \in \mathbb{R}^p$ represents the feature vector for observation j , and y_{jk} represents its corresponding binary indicator for clinical trajectory k . The clinical trajectories are not required to be mutually exclusive; therefore, each observation may contribute simultaneously to multiple trajectory contexts.

For a query observation i , CERL generates a trajectory-specific contextual representation by applying the representation function f_θ to $\mathbf{x}_i$ conditional on each contextual dictionary,

$$\mathbf{h}_i^{(k)} = f_\theta(\mathbf{x}_i \mid \mathcal{D}_k), \quad \mathbf{h}_i^{(k)} \in \mathbb{R}^d.$$

where f_θ was implemented using the Tabular Prior-data Fitted Network (TabPFN) (ver. 8.4.0)¹. Importantly, y_{ik} is not directly supplied as an input when constructing the representation of a held-out query observation. Instead, the outcome information contained in $\mathcal{D}_k$ provides the contextual information under which $\mathbf{x}_i$ is represented. Thus, each $\mathbf{h}_i^{(k)}$ represents the same observed feature vector under a different clinical trajectory context.

The trajectory-specific representations are subsequently concatenated within each observation to construct the final CERL representation:

$$\mathbf{z}_i^{\text{CERL}} = \text{concat}(\mathbf{h}_i^{(1)}, \mathbf{h}_i^{(2)}, \dots, \mathbf{h}_i^{(K)}) = [\mathbf{h}_i^{(1)} \parallel \mathbf{h}_i^{(2)} \parallel \dots \parallel \mathbf{h}_i^{(K)}],$$

such that

$$\mathbf{z}_i^{\text{CERL}} \in \mathbb{R}^{Kd}.$$

CERL therefore represents each observation using a unified embedding composed of multiple complementary clinical trajectory contexts. The resulting representation is continuous and does not intrinsically assign observations to discrete endotypes. When categorical clinical states are desired, an unsupervised clustering algorithm can subsequently be applied to $\mathbf{Z}^{\text{CERL}}$.

Synthetic data generation

Synthetic datasets were generated to assess the ability of CERL to recover known latent states under varying sample sizes, numbers of states, and linear or nonlinear state-generating mechanisms. Each dataset contained $N \in \{1,000, 1,500, 3,000\}$ observations and 20 continuous features. The number of ground-truth latent states was varied as $K \in \{2,3,4\}$. Each experimental configuration was independently repeated 50 times.

Generation of correlated input features: For each simulation replicate, a random matrix $\mathbf{A} \in \mathbb{R}^{20 \times 20}$ was generated from a standard normal distribution. To generate a sparse dependence structure, 90% of randomly selected elements of $\mathbf{A}$ were set to zero. The feature covariance matrix was then constructed as:

$$\mathbf{\Sigma} = \mathbf{A}\mathbf{A}^T + 10^{-3}\mathbf{I},$$

where the diagonal regularization term ensured positive definiteness. Input features were sampled according to:

$$\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$$

Thus, all observations shared the same marginal feature-generating distribution, while latent states were subsequently defined through multivariate functions of the observed features rather than through explicitly state-specific feature distributions.

Generation of latent states: A continuous latent score was first generated as a randomly weighted linear function of the input features,

$$z_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta}.$$

where β_0 is an intercept, i is an observation, and $\boldsymbol{\beta}$ is a vector of randomly generated feature coefficients.

For nonlinear simulation settings, the latent score was augmented by an interaction between two randomly selected features and rectified-linear transformations of three randomly selected features,

$$z_i = \beta_0 + \mathbf{x}_i^T \boldsymbol{\beta} + x_{ij}x_{ik} + \sum_{m \in \mathcal{S}} \text{ReLU}(x_{im}),$$

where j and k denote randomly selected interacting features and $\mathcal{S}$ contains three randomly selected features.

Generation of endotypes

Endotype prevalences were independently sampled for each simulation from a symmetric Dirichlet distribution with concentration parameter 5. Cumulative state prevalences were then used as quantile thresholds of the continuous latent score z to assign each observation to one of K ground-truth endotypes.

Generation of multiple clinical outcomes: Five binary outcomes were generated conditional on the ground-truth endotype states. For each endotype, baseline outcome probabilities were distributed between 0.05 and 0.70 and modified by outcome-specific effects ranging from 0.8 to 1.2. Random Gaussian perturbations with standard deviation 0.08 were added to the resulting state–outcome probability matrix, after which probabilities were constrained to the interval $[0.05, 0.70]$.

For observation i belonging to latent endotype c_i , each outcome was sampled as

$$Y_{ik} \sim \text{Bernoulli}(p_{c_ik}),$$

where p_{c_ik} represents the probability of outcome k for the corresponding latent state. Consequently, individual outcomes represented noisy manifestations of the underlying endotypes rather than deterministic endotype states, while dependence among outcomes arose through their shared association with the latent state.

Synthetic data experiments

We conducted synthetic data experiments to evaluate whether CERL could recover predefined endotypes under not only controlled linear, but also nonlinear data-generating mechanisms. Simulated datasets comprised 20 correlated continuous features, from which ground-truth endotypes were defined through an underlying continuous function of the input features. In nonlinear experiments, the latent function additionally incorporated feature interactions and nonlinear transformations. Multiple binary clinical outcomes were subsequently generated with prespecified endotype-dependent probabilities, providing partially overlapping outcome information associated with the underlying latent states. We evaluated sample sizes of 1,000, 1,500, and 3,000, with two to four latent states. In each scenario, 50 independent simulation replicates were constructed.

CERL representations were learned using the multiple simulated outcomes as contextual information and evaluated against conventional unsupervised approaches, including K-means clustering, Bayesian Gaussian mixture modeling, agglomerative clustering, and HDBSCAN, as well as PLS-DA^{2,3}-based outcome-informed dimensionality reduction followed by clustering. Recovery of the predefined endotypes was quantified using the adjusted Rand index (ARI). Generalizability of the CERL representation was additionally evaluated in a held-out test set comprising 33% of observations. CERL representations for the test set were generated using only the

development-set features and outcomes as contextual information, and K-means clustering was fitted exclusively on the development-set representations before being applied to the held-out test set.

All machine-learning analyses were performed in Python (ver. 3.11). Unsupervised machine learning methods were implemented using scikit-learn (ver. 1.9.0)⁴.

Preprocessing of eICU cohort data

To evaluate CERL in a real-world setting, we used a multicenter cohort of patients from the eICU Collaborative Research Database (eICU-CRD)⁵. The eICU-CRD contains 200,859 ICU encounters from 139,367 unique patients across multiple U.S. hospitals. For practical application, patients with sepsis were identified according to the Sepsis-3 criteria⁶, based on suspected infection and Sequential Organ Failure Assessment (SOFA) score of at least 2 points.

Because clinical deterioration related to sepsis may begin before ICU admission, we did not restrict sepsis ascertainment to measurements obtained after ICU entry. Instead, we defined an observation window extending from 7 days before ICU admission through ICU discharge. Suspected infection was identified by the collection of a microbiological culture within this observation window, and SOFA scores were reconstructed longitudinally from available clinical measurements. This approach was intended to capture sepsis manifestations preceding ICU admission among patients who subsequently required critical care.

To reconstruct daily SOFA scores, we first aggregated longitudinal measures within each day. For measurements with multiple values available on the same day, the value representing the worse condition was retained. The definition of worse case in each measurement is provided in Supplementary Table 1. Some measurements required for individual SOFA domains were not available at each time point. Treating every unmeasured component as missing would prevent calculation of the total SOFA score for a substantial proportion of observations. We therefore assigned a score of zero to a SOFA domain when the information required to identify organ dysfunction in that domain was not available. This operational assumption treats an unmeasured SOFA component as having no documented evidence of organ dysfunction rather than imputing an unobserved physiological measurement. SOFA domain scores were subsequently summed to obtain the total SOFA score used for cohort eligibility.

Application of these criteria identified 2,368 eligible ICU encounters among 2,127 unique patients. For patients with multiple eligible ICU encounters, only the earliest eligible encounter was retained. This excluded 241 subsequent encounters and yielded a final analytic cohort of 2,127 ICU encounters from 2,127 unique patients.

Clinical trajectory definition

Reconstructed daily SOFA scores were used to characterize longitudinal clinical trajectories among patients with sepsis. To incorporate mortality into the longitudinal

representation, SOFA scores were assigned a value of 25 from the time of death onward, representing a state beyond the maximum conventional SOFA score.

For trajectory clustering, daily SOFA scores were arranged in wide format, with each patient represented by a single row and each ordered time point relative to ICU admission represented by a separate feature. Thus, the temporal ordering of SOFA measurements was preserved within the feature representation. K-means clustering with $K = 10$ was then applied to these longitudinal SOFA vectors to identify data-driven groups with similar temporal, time-sensitive patterns.

The 10 data-driven trajectory clusters were subsequently reviewed and manually consolidated into four broader, clinically interpretable trajectory groups based on the observed shapes of their longitudinal SOFA profiles and mortality patterns. These groups were designated acute recovery, late recovery, acute mortality, and late mortality. The four groups were therefore not defined a priori using fixed SOFA thresholds or prespecified rates of change; rather, the acute and late designations reflected the temporal characteristics observed in the data-driven trajectory clusters. The resulting four trajectory groups were used as clinical contexts for CERL representation learning. The characteristic patterns observed in each trajectory group were as follows:

- Acute recovery: SOFA scores generally peaked around ICU admission and rapidly declined to < 2 within approximately three days after ICU admission.
- Late recovery: Persistent organ dysfunction with SOFA scores generally remaining > 2 during the first week after ICU admission and gradually declining toward < 2 by approximately 7 to 20 days after ICU admission.
- Acute death: Rapid progression to death, generally occurring within the first 10 days after ICU admission.
- Late death: Persistent organ dysfunction followed by death later in the clinical course, generally occurring between approximately 10 and 21 days after ICU admission.

These characteristics describe the empirical patterns observed in the data-driven trajectory clusters and were not used as prespecified numerical criteria for trajectory assignment.

Biological data preprocessing

To characterize biological states using routinely collected clinical data, we selected laboratory measurements from complete blood count and comprehensive metabolic panels. Lipid measurements were excluded because of substantial missingness ($>88\%$), whereas the retained laboratory measurements had missingness of approximately 1%–52%. The final feature set included red blood cell count, white blood cell count, hemoglobin, hematocrit, platelet count, glucose, calcium, sodium, potassium, bicarbonate, chloride, albumin, total protein, alanine aminotransferase, aspartate aminotransferase, total bilirubin, blood urea nitrogen, and creatinine. Missing values were imputed using the scikit-learn `IterativeImputer`⁴ with a maximum of 10 iterations. Features were subsequently standardized and used as inputs for CERL representation

construction. Longitudinal time stamps were not explicitly incorporated into either the imputation procedure or CERL representation construction.

Reproducibility assessment of CERL

To evaluate the reproducibility and clinical utility of CERL representations, we used a multicenter cohort of patients meeting Sepsis-3 criteria⁶ within the eICU Collaborative Research Database⁵. Eighteen routinely collected laboratory measurements, including components of the complete blood count and comprehensive metabolic panel, were used as input features. We defined six potentially overlapping clinical trajectory contexts for CERL representation learning. Four trajectories, including acute recovery, late recovery, acute mortality, and late mortality, were defined using the temporal course of Sequential Organ Failure Assessment (SOFA) scores and mortality. Invasive mechanical ventilation and renal replacement therapy during the ICU stay were included as two additional binary clinical contexts. These contexts across six binary indicators were used to construct context-specific CERL dictionaries. Detailed cohort selection and clinical trajectory definitions are provided above.

We conducted two institution-held-out validation experiments to evaluate the reproducibility of CERL representations and derived clinical states across healthcare institutions in the eICU. The eICU cohort included patients from 20 participating institutions. In each experiment, institutions were assigned to non-overlapping derivation and validation cohorts such that no institution contributed observations to both cohorts.

In the first validation experiment, institutions with more than 100 sepsis patients were assigned to the derivation cohort ($n = 9$ institutions; 1802 unique patients; 13,924 longitudinal observations), whereas the remaining institutions were defined as the validation cohort ($n = 11$ institutions; 325 unique patients; 2,955 longitudinal observations). In the second validation experiment, an alternative institution-level partition was used to assess robustness to cohort composition. Institutions with ≤ 100 or > 120 sepsis patients were assigned to the derivation cohort ($n = 17$ institutions; 1,784 unique patients; 14,384 longitudinal observations), whereas the remaining institutions were defined as the validation cohort ($n = 3$ institutions; 343 unique patients; 2493 longitudinal observations).

For each validation experiment, the CERL representation was constructed using the derivation cohort, and K-means clustering was fitted to the resulting representation to define three discrete CERL states (S0, S1, and S2). The fitted CERL representation model was subsequently applied to observations from the institution-held-out validation cohort to generate CERL embeddings, and the fitted K-means model was used to assign each validation observation to the CERL states.

Reproducibility was evaluated at three levels. First, representation-space reproducibility was assessed by calculating the mean CERL representation vector for each state separately in the derivation and validation cohorts. Cosine similarity between the corresponding state-specific mean representation vectors was used to quantify the directionality of the CERL state space across institutionally independent cohorts.

Second, biological reproducibility was assessed using the 18 laboratory measurements provided as biological inputs to CERL. State-specific mean laboratory profiles were calculated separately in the derivation and validation cohorts, and agreement between the corresponding profiles was quantified using Pearson correlation. This analysis evaluated whether CERL states retained similar underlying biological and physiological characteristics when transferred to patients from held-out institutions.

Third, clinical reproducibility was evaluated by estimating the proportion of each prespecified clinical trajectory within each CERL state. These included acute recovery, late recovery, acute mortality, late mortality, invasive mechanical ventilation, and renal replacement therapy. State-specific outcome proportions were calculated independently in the derivation and validation cohorts and compared to assess whether the clinical characteristics associated with each CERL state were preserved across institutions.

Because individual patients contributed multiple longitudinal observations, patient-level block bootstrap resampling was used to estimate uncertainty while preserving within-patient dependence. Reproducibility statistics and state-specific outcome proportions were recalculated for each bootstrap replicate, and 95% confidence intervals were obtained from the resulting bootstrap distributions.

Benchmarking with conventional unsupervised machine learning models

CERL was designed as a biological and clinical state representation framework in which multiple, potentially overlapping clinical trajectories provide contextual information during representation learning. To evaluate whether this contextual representation provides additional information beyond conventional outcome-agnostic phenotyping, we benchmarked CERL against K-means clustering and a Bayesian Gaussian mixture model (BGM) applied directly to the input laboratory measurements. Comparisons focused on three complementary properties: state reproducibility, biological reproducibility, and clinical trajectory informativeness.

State and biological reproducibility were evaluated using the same institution-held-out validation framework described above. For each comparator, clustering models were fitted in the derivation cohort and subsequently applied to the held-out validation cohort without refitting. State reproducibility was assessed by comparing state-specific profiles between the derivation and validation cohorts, while biological reproducibility was quantified by the agreement of state-specific mean laboratory profiles across cohorts.

We further evaluated whether the derived state partitions retained information associated with each of the six prespecified clinical trajectories. State membership was encoded as a categorical explanatory variable, and a separate binary logistic regression model was fitted for each clinical trajectory. Clinical trajectory discrimination was quantified using the area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC). These metrics summarize discrimination across a range of probability thresholds and therefore allowed us to assess the clinical informativeness of the state partitions without selecting a specific classification threshold. In contrast, threshold-dependent classification metrics, including

accuracy, F1 score, sensitivity, specificity, positive predictive value, and negative predictive value, were not evaluated. The derived states were not constructed to correspond directly to individual clinical trajectories, and the logistic regression models were used to quantify trajectory discrimination associated with state membership rather than to develop or evaluate clinical prediction models.

Because multiple laboratory observations were available from individual patients, patient-level block bootstrap resampling was used to compare CERL with conventional clustering while preserving within-patient dependence. For each bootstrap replicate, the same resampled patients were used for both approaches, and differences in AUROC and AUPRC (Δ AUROC and Δ AUPRC) were calculated directly between CERL and each comparator. The resulting paired bootstrap distributions were used to derive 95% confidence intervals.

To evaluate whether CERL captured clinical trajectory-related organization beyond conventional outcome-agnostic phenotyping, K-means clustering was applied directly to the laboratory measurements used as CERL inputs. The number of clusters was selected using the silhouette score, resulting in four laboratory-defined states.

Within each laboratory-defined state, pairwise distances were calculated between observations belonging to the same versus different clinical trajectories. Clinical trajectories included acute recovery, late recovery, acute death, and late death. Trajectory-related organization was quantified as the ratio of the median between-trajectory distance to the median within-trajectory distance. A ratio greater than 1 indicated greater separation between distinct clinical trajectories. Ratios were compared between the CERL-derived and laboratory-derived PHATE spaces to assess whether CERL captured additional trajectory-related organization within laboratory-defined biological states.

All machine-learning analyses were performed in Python (ver. 3.11). Unsupervised machine learning methods were implemented using scikit-learn (ver. 1.9.0)⁴. Statistical analyses and figure generation were performed using R (version 4.5.3).

References

1. Hollmann N, Müller S, Purucker L, et al. Accurate predictions on small data with a tabular foundation model. *Nature*. 2025;637(8045):319-326. doi:10.1038/s41586-024-08328-6
2. Barker M, Rayens W. Partial least squares for discrimination. *Journal of Chemometrics*. 2003;17(3):166-173. doi:10.1002/cem.785
3. Ståhle L, Wold S. Partial least squares analysis with cross-validation for the two-class problem: A Monte Carlo study. *Journal of Chemometrics*. 1987;1(3):185-196. doi:10.1002/cem.1180010306
4. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. Published online 2012. doi:10.48550/ARXIV.1201.0490
5. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data*. 2018;5(1):180178. doi:10.1038/sdata.2018.178
6. 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. doi:10.1001/jama.2016.0287

Supplementary Tables

Supplementary Table 1. Definitions of daily worst values used to calculate SOFA scores

Respiratory		409
P/F ratio	Min	410
Coagulation		411
Platelets	Min	412
Liver		413
Total bilirubin	Max	414
Cardiovascular		415
Mean arterial pressure	Min	416
Vasoactive drug dose	Max	417
Central nervous system		418
Glasgow Coma Scale	Min	419
Renal		420
Creatinine	Max	421
		422
		423
		424

450 Supplementary Table 2. Demographic characteristics of patients meeting Sepsis-3 criteria in the eICU Collaborative
451 Research Database, stratified by clinical trajectory

	Sepsis Population n = 2127	Acute Death n = 158	Acute Recovery n = 1638	Late Death n = 38	Late recovery n = 293	RRT n = 80	Invasive Ventilation n = 321
Male, n (%)	1189 (55.90)	86 (54.43)	905 (55.25)	23 (60.53)	175 (59.73)	47 (58.75)	177 (55.14)
Female, n (%)	938 (44.10)	72 (45.57)	733 (44.75)	15 (39.47)	118 (40.27)	33 (41.25)	144 (44.86)
Age, mean (SD), y	64.52 (16.02)	67.58 (16.19)	64.57 (16.18)	65.42 (13.80)	62.44 (14.99)	62.74 (13.51)	61.69 (16.17)
Race and ethnicity							
African American, n (%)	339 (15.94)	18 (11.39)	256 (15.63)	6 (15.79)	59 (20.14)	35 (43.75)	40 (12.46)
Asian, n (%)	8 (0.38)	1 (0.63)	5 (0.31)	0 (0.00)	2 (0.68)	0 (0.00)	1 (0.31)
Caucasian, n (%)	1701 (79.97)	131 (82.91)	1319 (80.53)	29 (76.32)	222 (75.77)	41 (51.25)	266 (82.87)
Hispanic, n (%)	21 (0.99)	1 (0.63)	15 (0.92)	1 (1.37)	4 (1.37)	3 (3.75)	3 (0.93)
Other/Unknown, n (%)	58 (2.73)	7 (4.43)	43 (2.63)	2 (2.05)	6 (2.05)	1 (1.25)	11 (3.43)
Septic Shock, n (%)	151 (7.10)	30 (18.99)	86 (5.25)	4 (10.53)	31 (10.58)	6 (7.50)	96 (29.91)
Medical or Surgical Admission, n (%)	1802 (84.72)	148 (93.67)	1378 (84.13)	35 (92.11)	241 (82.25)	74 (92.50)	286 (89.10)
Comorbidities							
Cardiovascular Disease, n (%)	1216 (57.17)	89 (56.33)	916 (55.92)	23 (60.53)	188 (64.16)	73 (91.25)	155 (48.29)
Hypertension, n (%)	988 (46.45)	68 (43.04)	745 (45.48)	18 (47.37)	157 (53.58)	58 (72.5)	123 (38.32)
Cardiac Arrhythmia, n (%)	337 (15.84)	21 (13.29)	263 (16.06)	6 (15.79)	47 (16.04)	16 (20.00)	33 (10.28)
Congestive Heart Failure, n (%)	306 (14.39)	18 (11.39)	241 (14.71)	4 (10.53)	43 (14.68)	21 (26.25)	36 (11.21)
Coronary Artery Disease, n (%)	309 (14.53)	18 (11.39)	232 (14.16)	7 (18.42)	52 (17.75)	22 (27.50)	56 (17.45)

Valvular Disease, n (%)	61 (2.87)	2 (1.27)	41 (2.50)	1 (2.63)	17 (5.80)	1 (1.25)	8 (2.49)
Peripheral Vascular Disorders, n (%)	141 (6.63)	11 (6.96)	102 (6.23)	4 (10.53)	24 (8.19)	12 (15.00)	18 (5.61)
Diabetes, n (%)	603 (28.35)	34 (21.52)	451 (27.53)	8 (21.05)	110 (37.54)	39 (48.75)	64 (19.94)
Pulmonary Disease, n (%)	528 (24.82)	39 (24.68)	401 (24.48)	14 (36.84)	74 (25.26)	21 (26.25)	92 (28.66)
Chronic Pulmonary Disease, n (%)	491 (23.08)	37 (23.42)	369 (22.53)	13 (34.21)	72 (24.57)	20 (25.00)	89 (27.73)
Pulmonary Circulation Disorders, n (%)	40 (1.88)	2 (1.27)	34 (2.08)	1 (2.63)	3 (1.02)	2 (2.50)	5 (1.56)
Respiratory Failure, n (%)	50 (2.35)	1 (0.63)	41 (2.50)	3 (7.89)	5 (1.71)	0 (0.00)	7 (2.18)
Renal Failure, n (%)	341 (16.03)	25 (15.82)	242 (14.77)	7 (18.42)	67 (22.87)	80 (100.00)	40 (12.46)
Metabolic, n (%)	220 (10.34)	15 (9.49)	161 (9.83)	5 (13.16)	39 (13.31)	12 (15.00)	28 (8.72)
Hypothyroidism, n (%)	210 (9.87)	13 (8.23)	154 (9.40)	5 (13.15)	38(12.97)	9 (11.25)	27 (8.41)
Hyperthyroidism, n (%)	10 (0.47)	2 (1.27)	7 (0.43)	0 (0.00)	1 (0.34)	3 (3.75)	1 (0.31)
Cancer, n (%)	184 (8.65)	17 (10.76)	138 (8.42)	4 (10.53)	25 (8.53)	7 (8.75)	19 (5.92)
Cerebrovascular Disease, n (%)	189 (8.89)	15 (9.49)	132 (8.06)	3 (7.89)	39 (13.31)	11 (13.75)	17 (5.30)
TIA	49 (2.30)	2 (1.27)	34 (2.08)	2 (5.26)	11 (3.75)	3 (3.75)	3 (0.93)
Stroke	159 (7.48)	14 (8.86)	111 (6.78)	2 (5.26)	32 (10.92)	9 (11.25)	17 (5.30)
Neuromuscular Disease, n (%)	17 (0.80)	0 (0.00)	12 (0.73)	0 (0.00)	5 (1.71)	1 (1.25)	1 (0.31)
Seizures, n (%)	137 (6.44)	10 (6.33)	108 (6.59)	4 (10.53)	15 (5.12)	7 (8.75)	33 (10.28)
Gastrointestinal Disorders, n (%)	95 (4.47)	11 (6.96)	65 (3.97)	0 (0.00)	19 (6.48)	7 (8.75)	10 (3.12)
Liver disease, n (%)	55 (2.59)	6 (3.80)	37 (2.26)	0 (0.00)	12 (4.1)	3 (3.75)	8 (2.49)
Peptic Ulcer Disease (no bleeding), n (%)	31 (1.46)	3 (1.90)	22 (1.34)	0 (0.00)	6 (2.05)	3 (3.75)	0 (0.00)

Peptic Ulcer Disease (with bleeding), n (%)	10 (0.47)	2 (1.27)	6 (0.37)	0 (0.00)	2 (0.68)	1 (1.25)	3 (0.93)
Dementia, n (%)	81 (3.81)	6 (3.80)	69 (4.21)	0 (0.00)	6 (2.05)	2 (2.50)	9 (2.80)
Rheumatoid and Collagen Diseases, n (%)	50 (2.35)	6 (3.80)	33 (2.01)	2 (5.26)	9 (3.07)	4 (5.00)	11 (3.43)
Coagulopathy, n (%)	13 (0.61)	0 (0.00)	10 (0.61)	0 (0.00)	3 (1.02)	1 (1.25)	0 (0.00)

*patients may have multiple conditions therefore prevalences are not expected to equal 100

Supplementary Figures

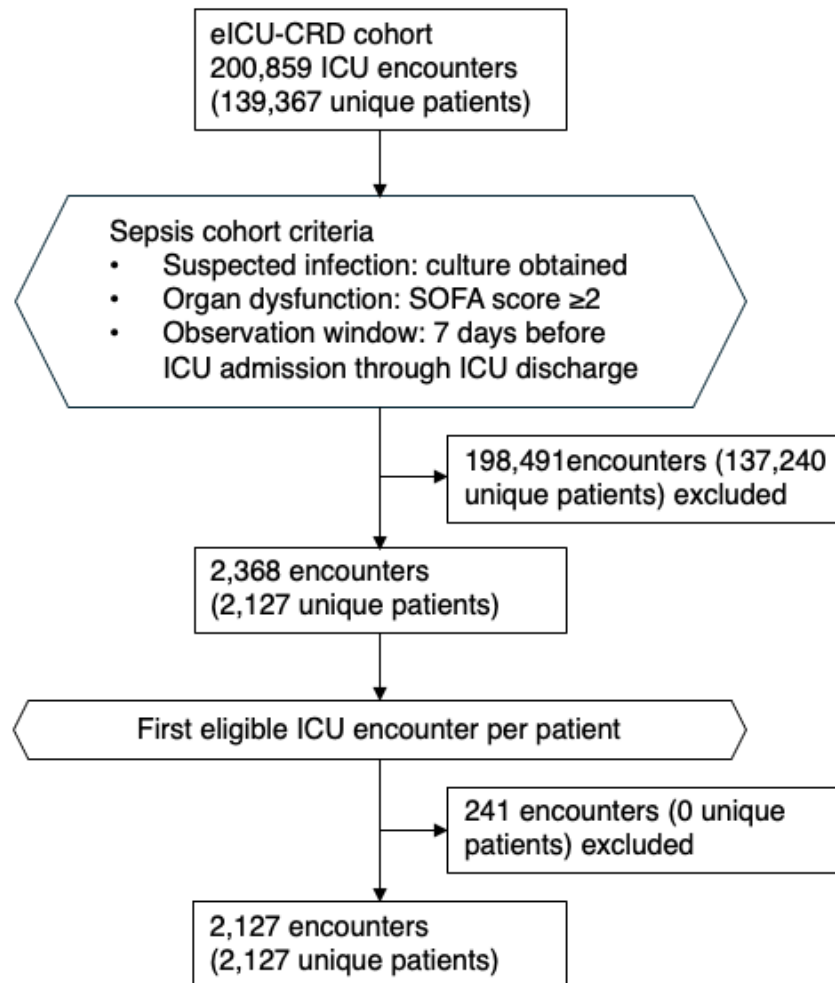


Figure S1. Flowchart for identification of patients with sepsis in the eICU Collaborative Research Database. Sepsis was identified using an operationalized Sepsis-3 definition based on suspected infection and a SOFA score of ≥ 2 points.

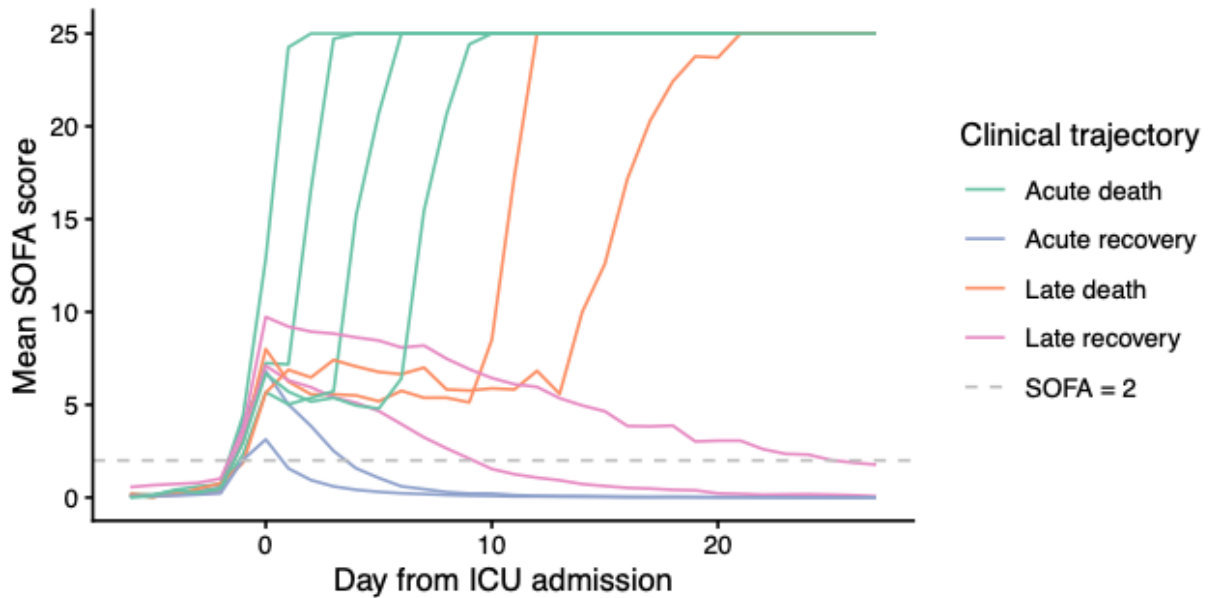
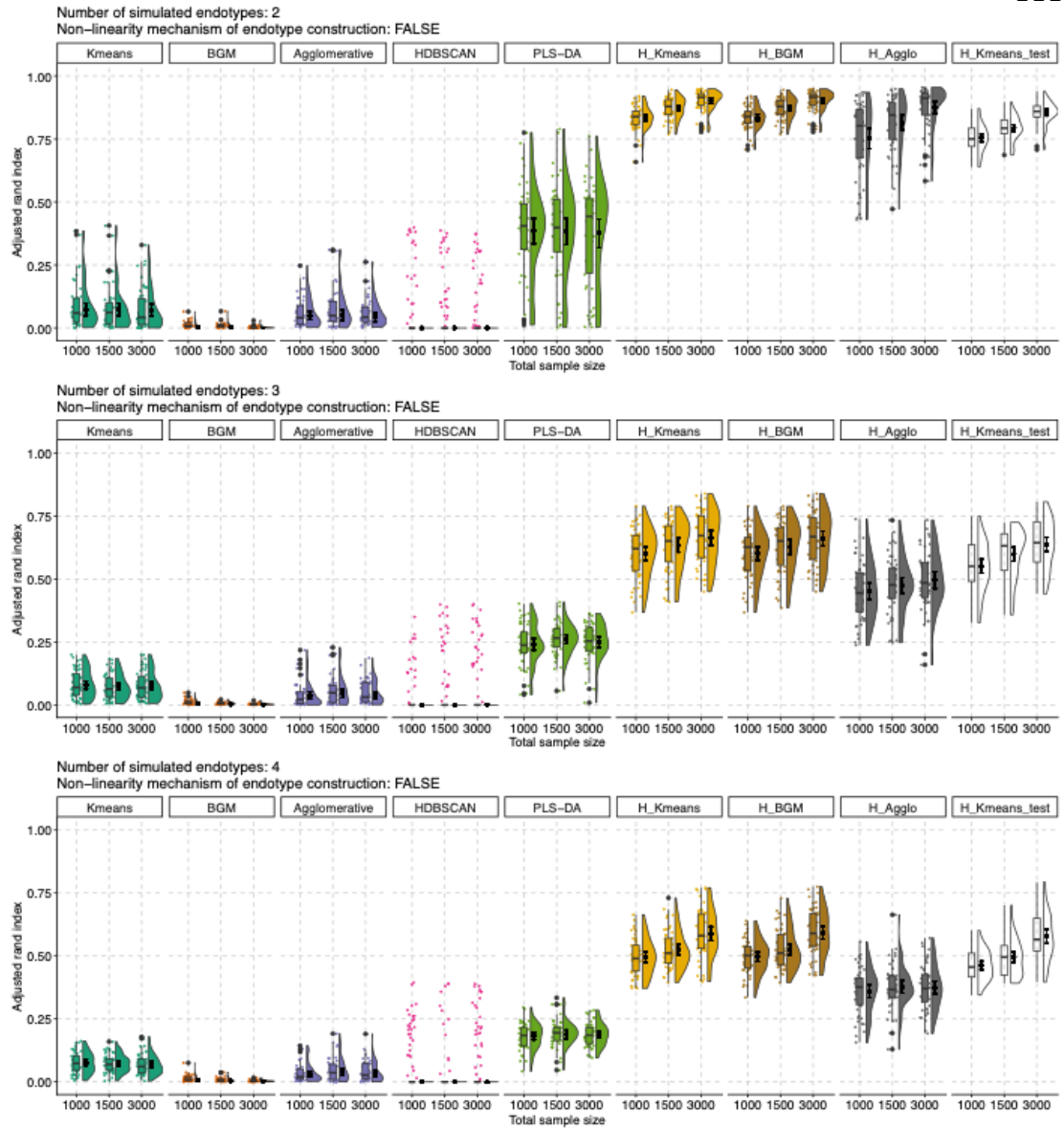
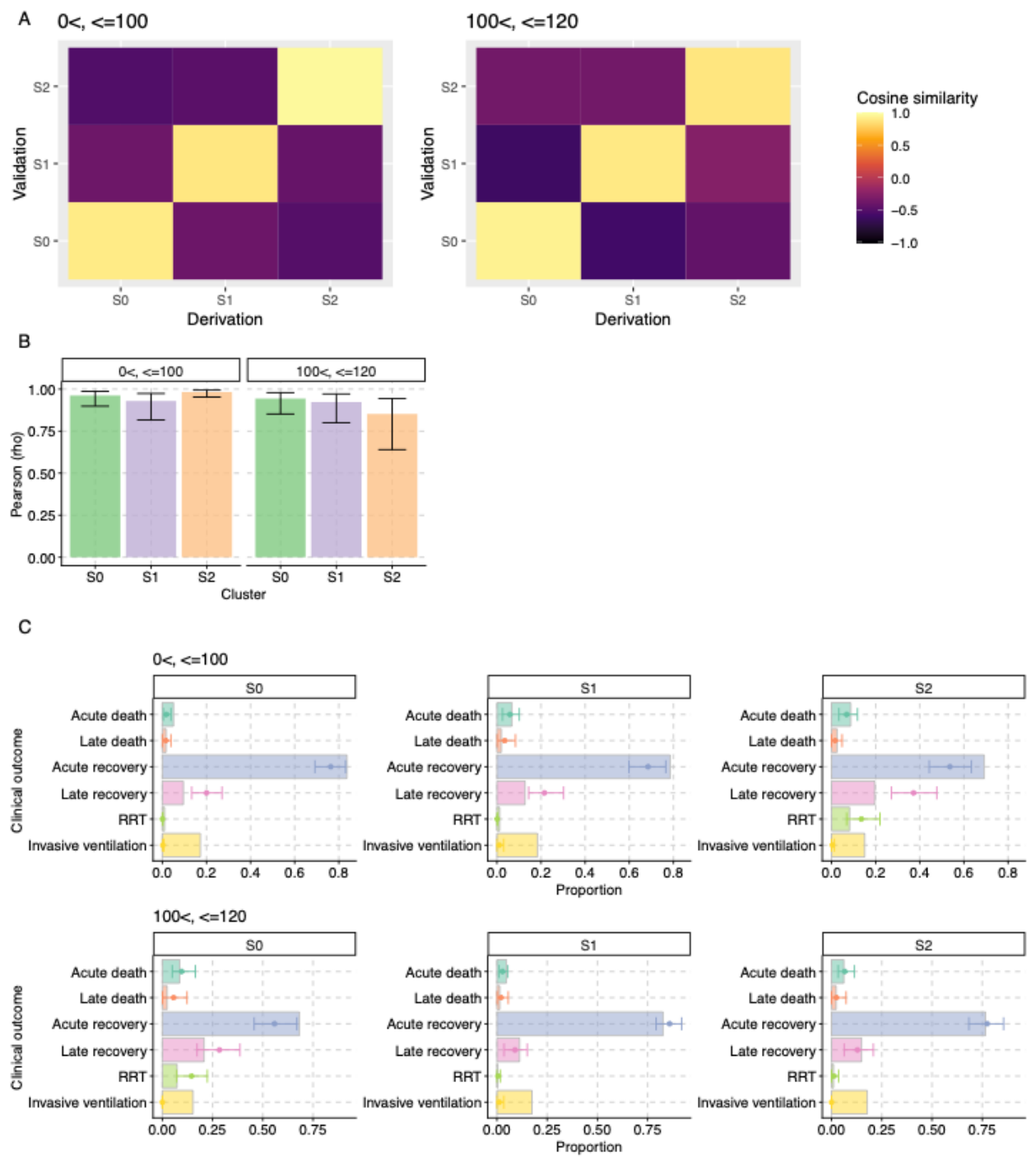


Figure S2. Daily Sequential Organ Failure Assessment (SOFA) scores were used to identify 10 initial groups based on their temporal patterns. These groups were subsequently manually consolidated into four clinically interpretable trajectory groups: acute recovery, late recovery, acute death, and late death. The dashed line indicates a SOFA score of 2 for reference.



586 Figure S3. Synthetic data evaluation of CERN for latent endotype recovery. Synthetic
587 datasets were generated from correlated multivariate normal distributions, and
588 predefined latent endotypes were constructed using linear functions of the simulated
589 features. Multiple partially overlapping clinical outcomes were subsequently generated
590 conditional on the predefined endotypes and used as clinical trajectory context for
591 CERN. Data were separated into development and held-out validation sets, with
592 validation representations generated using only the development-set context. Recovery
593 of the predefined ground-truth endotypes was quantified using the adjusted Rand index
594 (ARI). CERN was compared with conventional unsupervised clustering approaches and
595 outcome-informed dimensionality reduction across varying sample sizes and numbers
596 of predefined endotypes.
597

598



599
600
601
602
603
604
605

Figure S4. Institution-held-out validation of CERL in the eICU-CRD. Two institution-held-out validation experiments were performed to evaluate the reproducibility of CERL across independent institutional partitions. A) Cosine similarity of CERL representations between the derivation and validation cohorts across CERL states. B) Pearson correlations of laboratory profiles used to construct CERL representations between corresponding states in the derivation and validation cohorts. C) Clinical trajectory proportions across CERL states in the derivation cohort (bars) and validation cohort (points), with 95% confidence intervals estimated using patient-level bootstrap resampling.

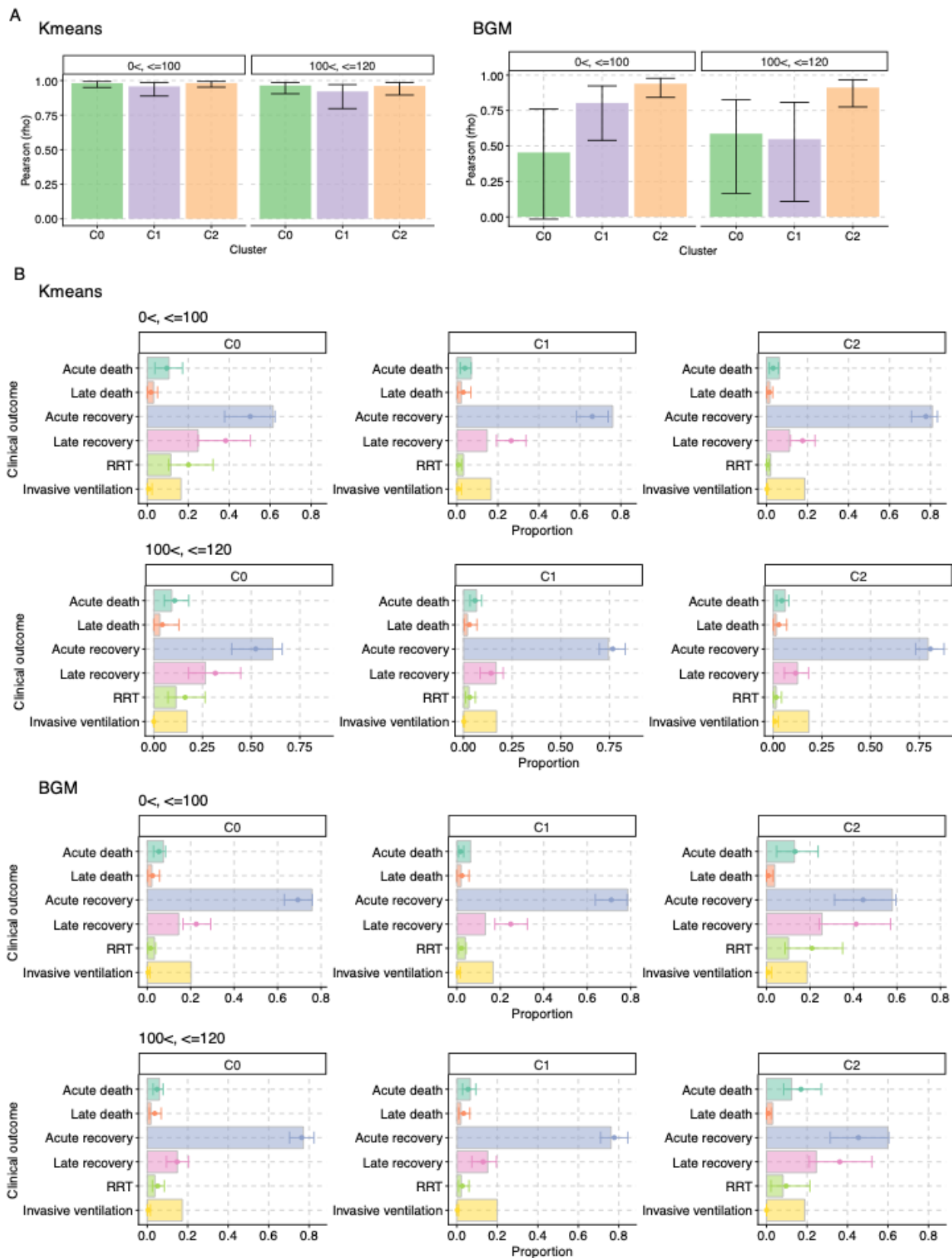


Figure S5. Benchmarking CERL against conventional unsupervised machine-learning approaches. CERL was compared with K-means clustering and Bayesian Gaussian mixture modeling applied directly to laboratory measurements. A) Pearson correlations of laboratory profiles between corresponding states in the derivation and validation cohorts for each approach. B) Clinical trajectory proportions across states in the derivation cohort (bars) and validation cohort (points) for each approach, with 95% confidence intervals estimated using patient-level bootstrap resampling.

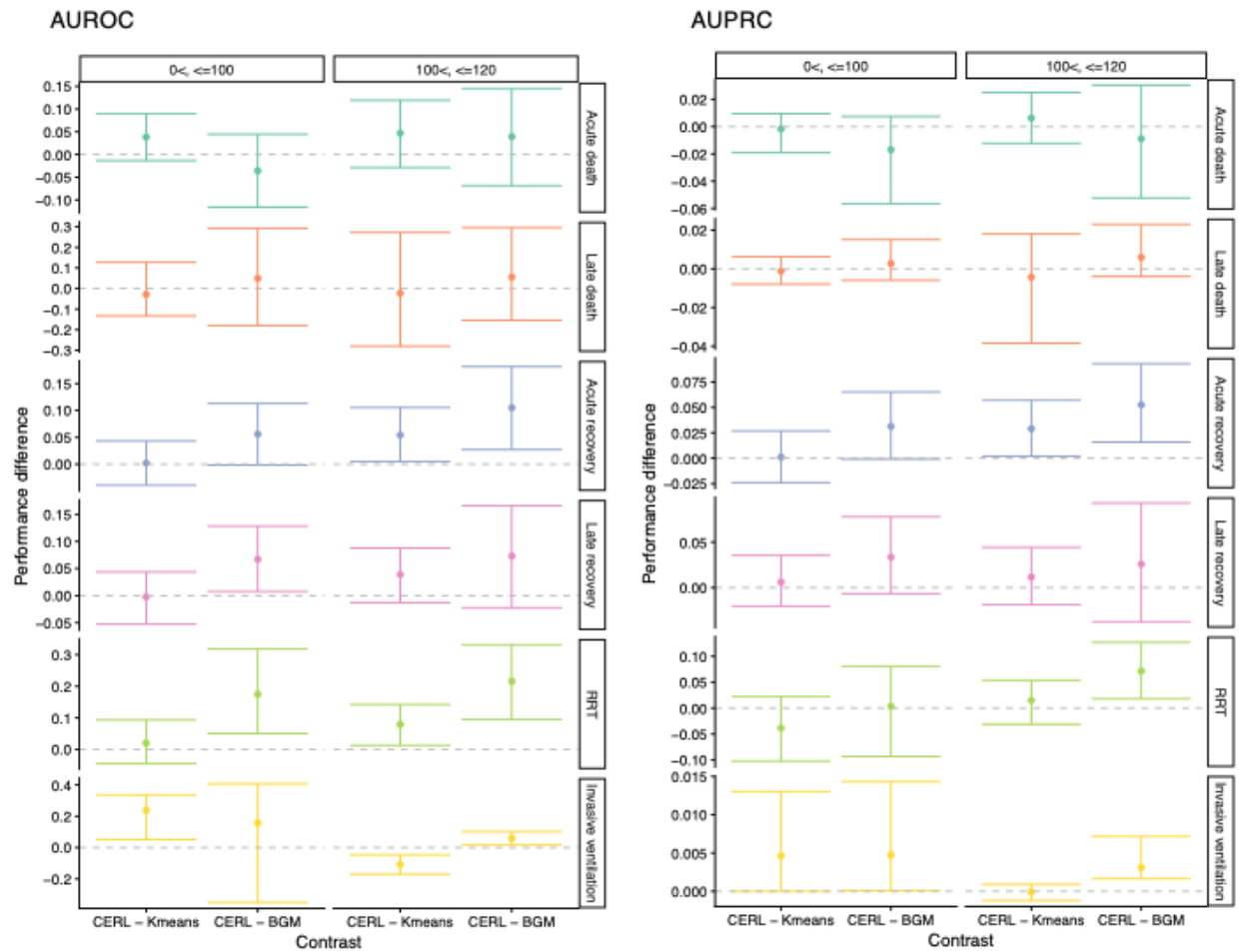
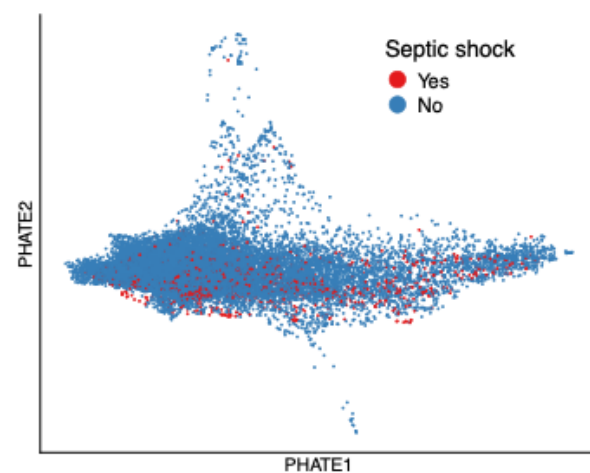
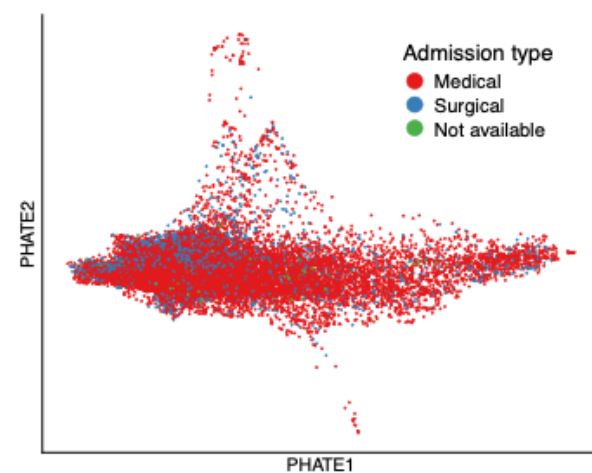
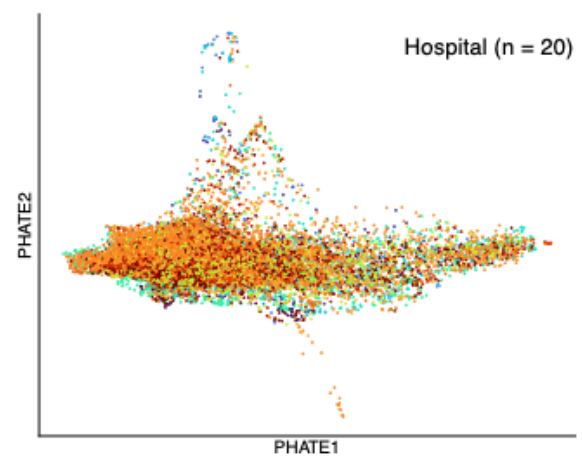
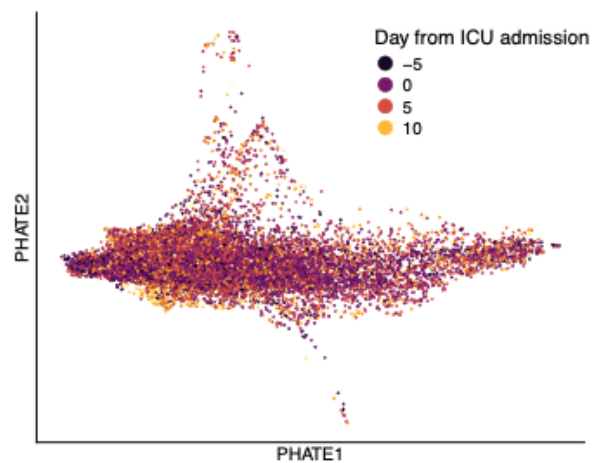
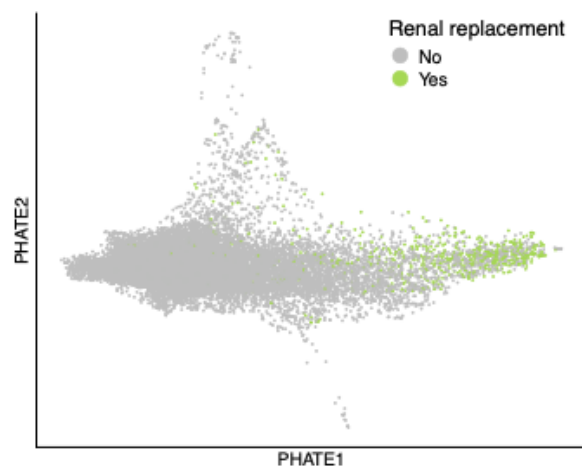
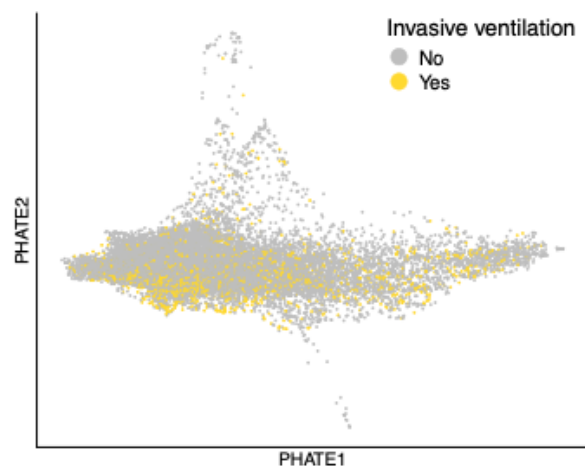


Figure S6. Benchmarking CERL against conventional unsupervised machine-learning approaches for clinical trajectory informativeness. Clinical trajectory informativeness of CERL-derived states was compared with states identified using conventional unsupervised approaches using the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). Results are presented as the difference in each metric between CERL and the comparator model (CERL – comparator). Positive values indicate greater trajectory informativeness for CERL, whereas negative values indicate greater informativeness for the comparator model. Error bars represent 95% confidence intervals estimated using patient-level bootstrap resampling.



688
689
690
691
692

Figure S7. Potential of Heat-diffusion for Affinity-based Transition Embedding (PHATE) projection of CERL representations, colored by invasive mechanical ventilation, renal replacement therapy, day from ICU admission, institution, admission type (medical or surgical), and presence of septic shock. These projections illustrate the distribution of clinical characteristics across the continuous CERL state space.

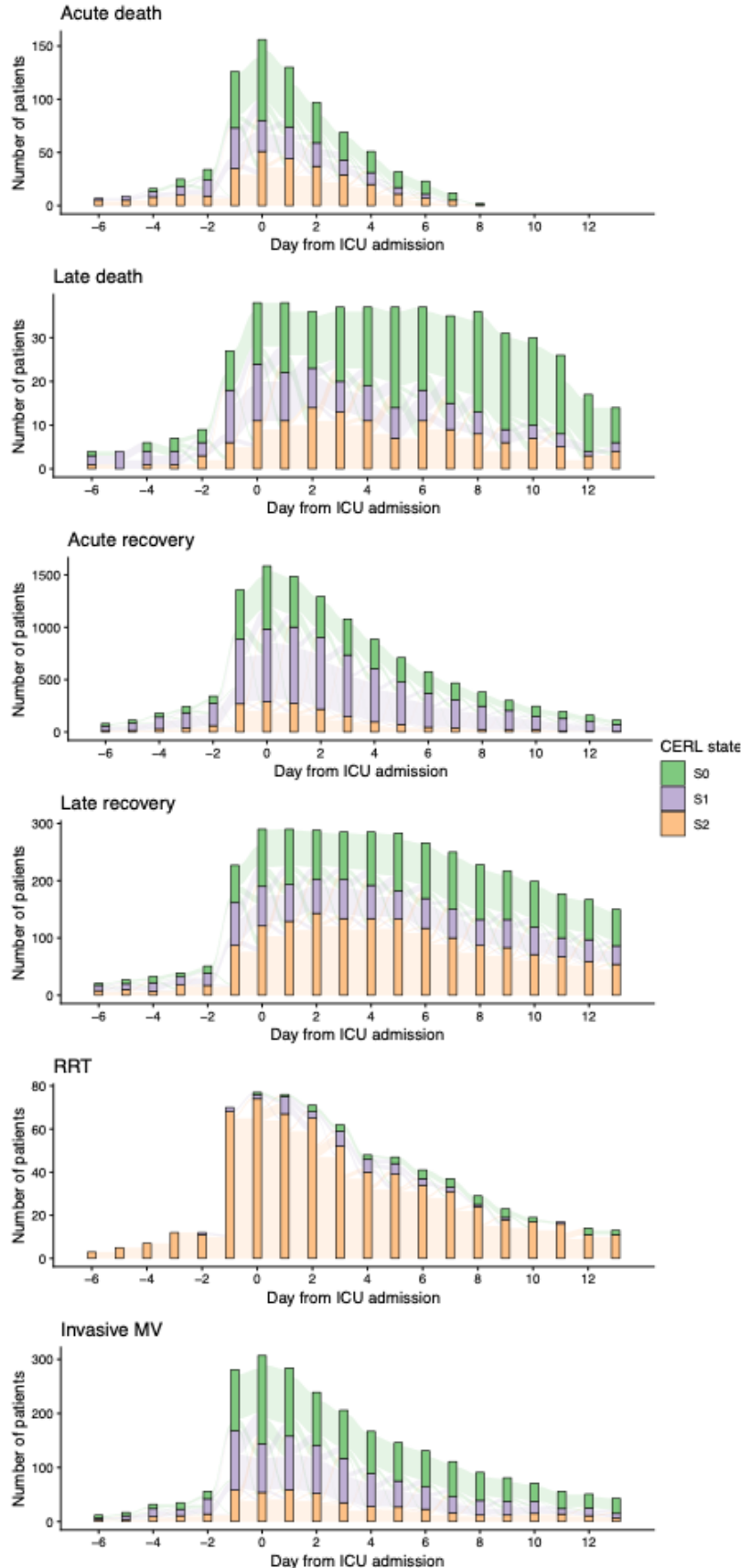


Figure S8. Temporal redistribution of CERL states across clinical trajectories. Alluvial plots illustrate changes in the distribution of CERL states over time across acute recovery, late recovery, acute death, and late death trajectories.

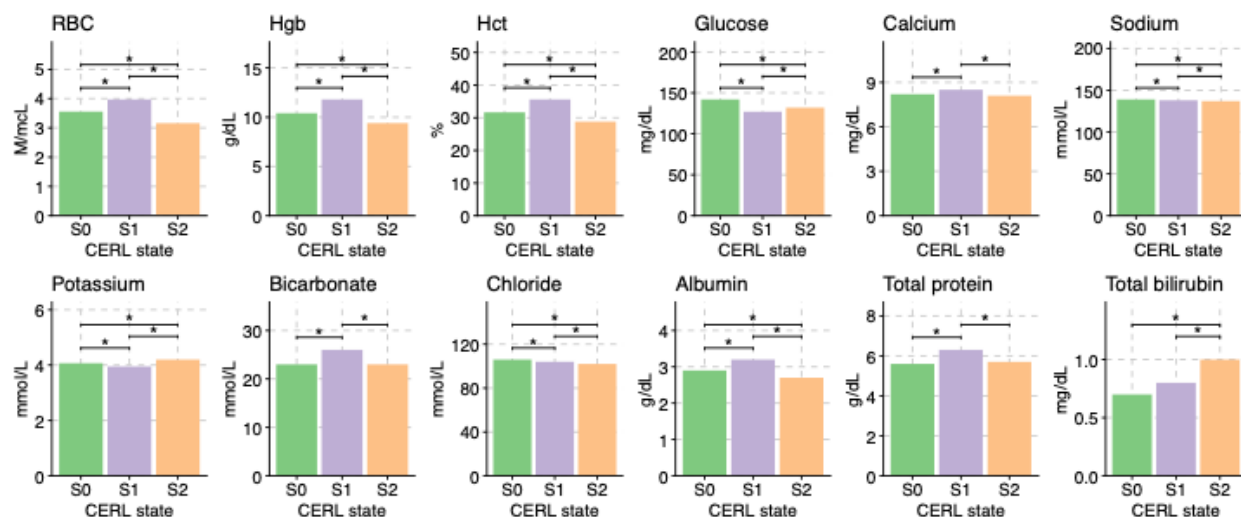


Figure S9. CERL state-specific laboratory profiles. Laboratory measurements obtained from one day before to one day after ICU admission were summarized by the mean value for each measurement. Differences across CERL states were evaluated using linear mixed-effects models with patient-specific random intercepts to account for repeated measurements. Pairwise comparisons between CERL states were performed with Tukey adjustment for multiple comparisons. Statistical significance was defined as an adjusted $P < 0.05$ and is indicated by an asterisk (*).

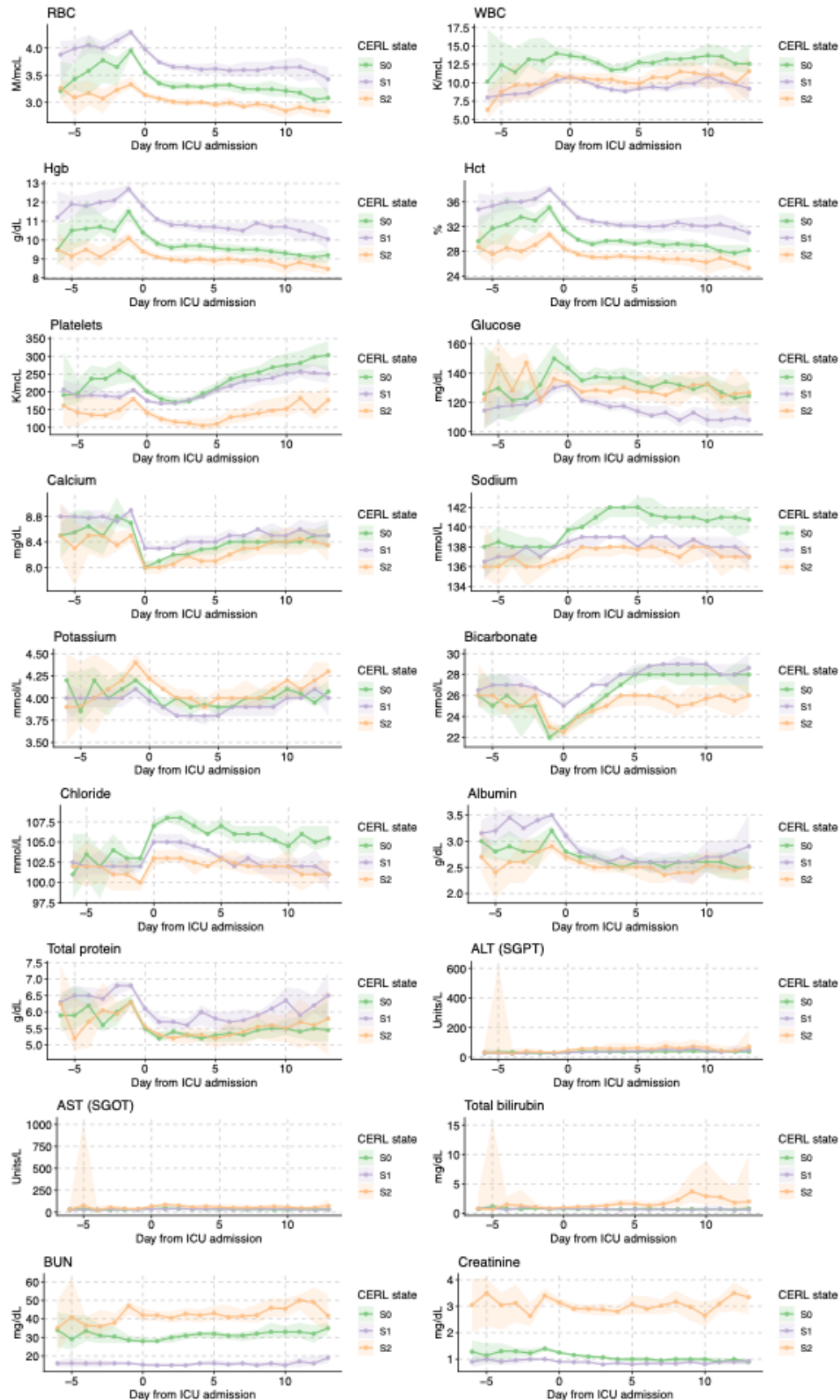


Figure S10. Longitudinal laboratory profiles across CERL-derived states. At each time point, the median value of each laboratory measurement was estimated for each CERL state. Shaded areas represent 95% confidence intervals estimated using bootstrap resampling for each time point.

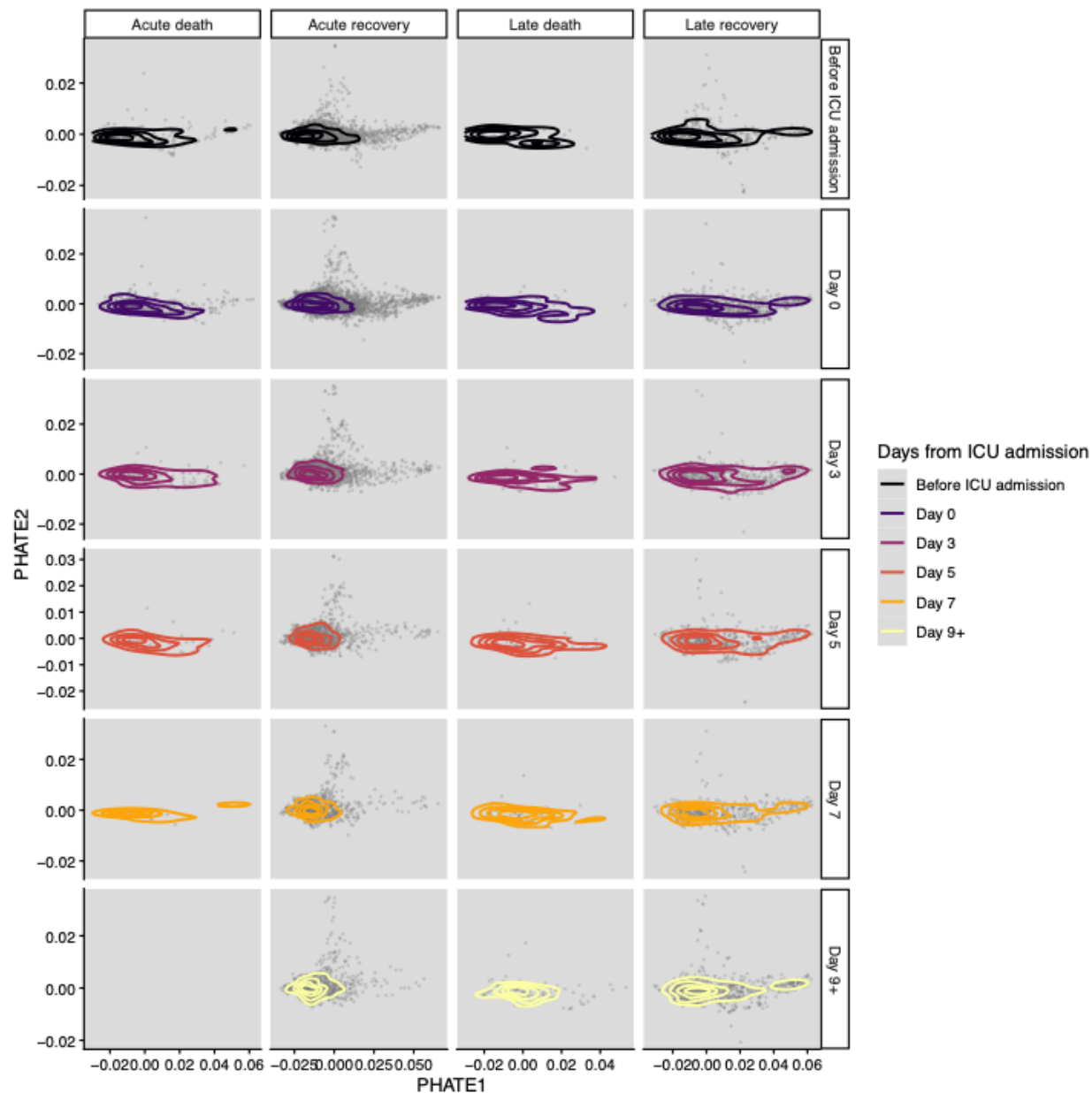
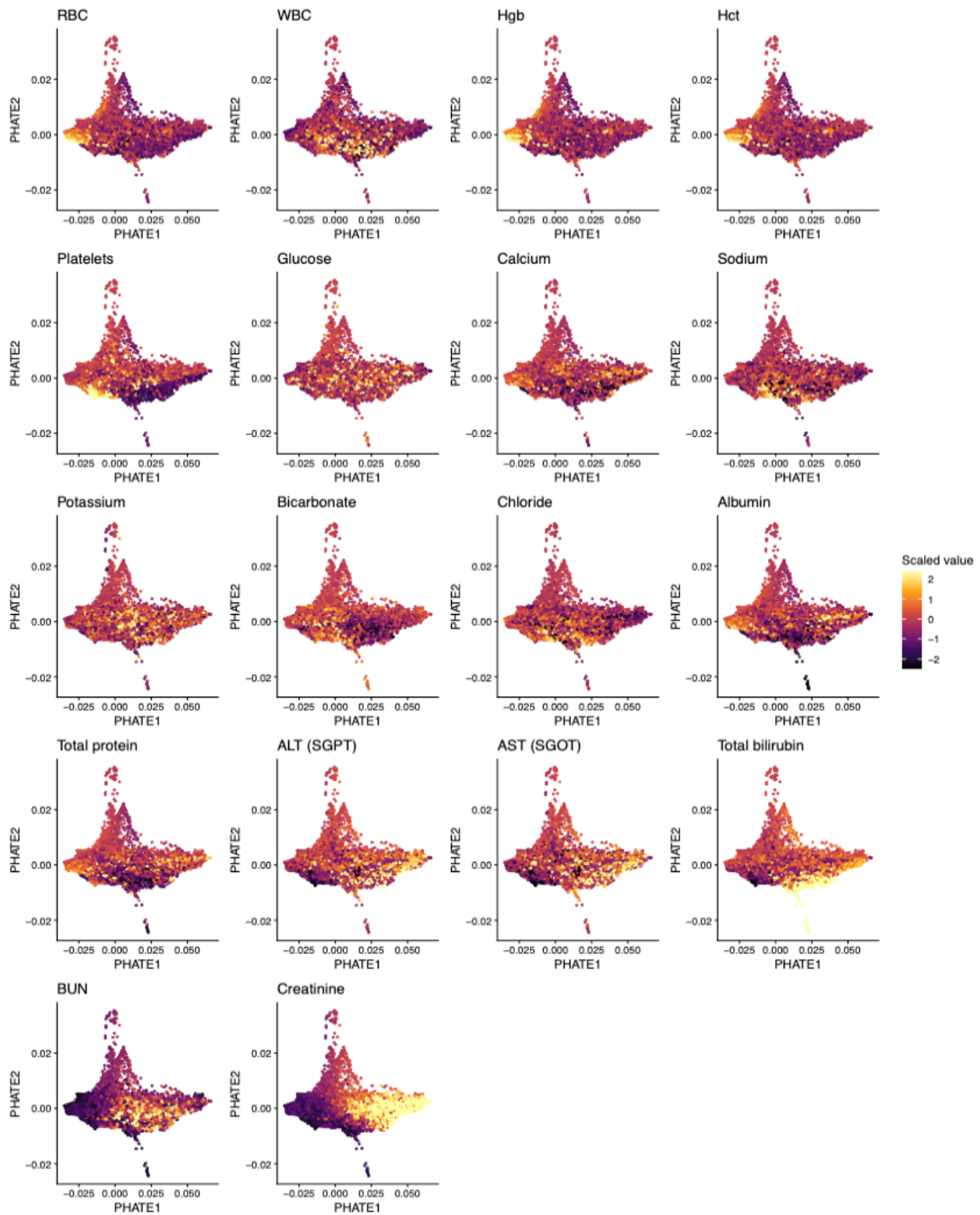


Figure S11. Temporal density distribution of clinical trajectories across the CERL state space. Density distributions of observations within the continuous CERL representation on PHATE projection were estimated at discrete time points for each clinical trajectory group. Density contours are colored according to time from ICU admission.



967
968
969
970

971 Figure S12. Spatial distribution of laboratory measurements on the continuous CERL
972 representation using PHATE projection. Laboratory measurements were standardized
973 and subsequently imputed as described in the Methods. For visualization, values below
974 the first quartile minus 1.5 times the interquartile range (IQR) or above the third quartile
975 plus 1.5 times the IQR were capped at the corresponding lower or upper threshold,
976 respectively, to reduce the influence of extreme values on color scaling. The resulting
977 values were used to color observations across the PHATE projection.

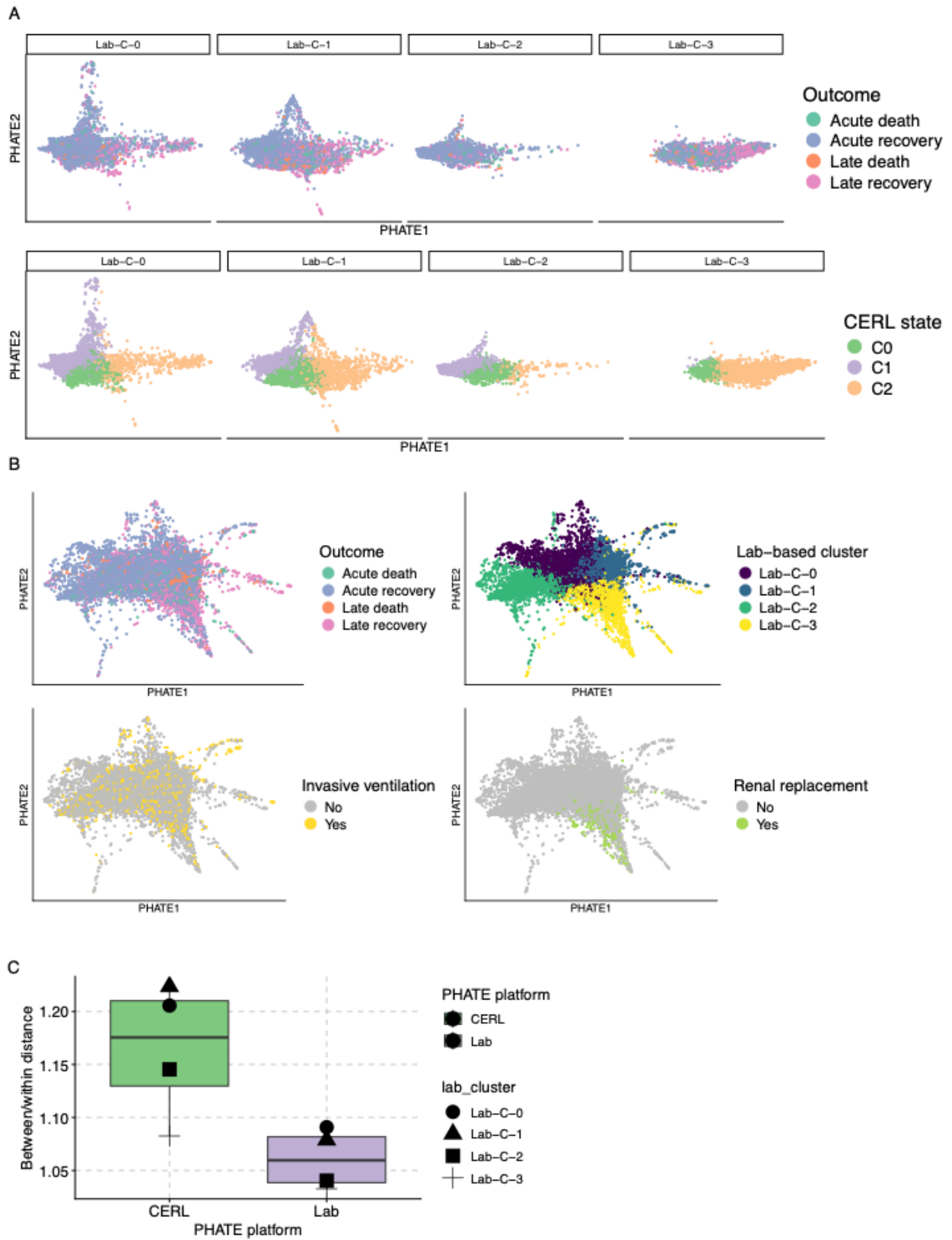


Figure S13. CERL captures additional clinical trajectory-related organization within laboratory-defined biological states. A) Observations stratified by four laboratory-defined states (Lab-C-0 to Lab-C-3) and visualized within the CERL-derived PHATE representation. Points are colored by clinical trajectory (top) or CERL state (bottom). B) Laboratory measurements were used to construct PHATE space, colored by clinical trajectory, laboratory-defined cluster, invasive mechanical ventilation, and renal replacement therapy. C) Ratio of between-trajectory to within-trajectory pairwise distances within each laboratory-defined state, calculated in the CERL-derived and laboratory-derived PHATE representations. Symbols indicate individual laboratory-defined clusters, and boxplots summarize the ratio distribution across lab-derived clusters.