PNAS RESEARCH ARTICLE | MEDICAL SCIENCES APPLIED MATHEMATICS

Robust extraction of pneumonia-associated clinical states from electronic health records

Feihong Xu[a,b], Félix L. Morales[a], and Luís A. Nunes Amaral[a,c,d,e,f,g,1]

Affiliations are included on p. 6.

Edited by Simon Levin, Princeton University, Princeton, NJ; received September 5, 2024; accepted September 30, 2024

Mining of electronic health records (EHR) promises to automate the identification of comprehensive disease phenotypes. However, the realization of this promise is hindered by the unavailability of generalizable ground-truth information, data incompleteness and heterogeneity, and the lack of generalization to multiple cohorts. We present here a data-driven approach to identify clinical states that we implement for 585 critical care patients with suspected pneumonia recruited by the SCRIPT study, which we compare to and integrate with 9,918 pneumonia patients from the MIMIC-IV dataset. We extract and curate from their structured EHRs a primary set of clinical features (53 and 59 features for SCRIPT and MIMIC-IV, respectively), including disease severity scores, vital signs, and so on, at various degrees of completeness. We aggregate irregular time series into daily frequency, resulting in 12,495 and 94,684 patient-day pairs for SCRIPT and MIMIC, respectively. We define a "common-sense" ground truth that we then use in a semisupervised pipeline to optimize choices for data preprocessing, and reduce the feature space to four principal components. We describe and validate an ensemble-based clustering method that enables us to robustly identify five clinical states, and use a Gaussian mixture model to quantify uncertainty in cluster assignment. Demonstrating the clinical relevance of the identified states, we find that three states are strongly associated with disease outcomes (dying vs. recovering), while the other two reflect disease etiology. The outcome associated clinical states provide significantly increased discrimination of mortality rates over standard approaches.

EHR mining | high dimensionality | clustering | multicenter integration

Significance

Pneumonia, a leading cause of death, is characterized by a variety of pathologies, symptoms, comorbidities, and clinical courses, which creates enormous challenges for treatment prognosis. Electronic health records offer the potential to identify granular clinical states within pneumonia but multimodality, missing data, and an unknown ground-truth have prevented the achievement of this goal. We report here on advances in i) the definition of common-sense ground-truths that enable us to decide upon feature selection, data imputation, and dimensionality reduction choices, ii) an ensemble density peak clustering approach that robustly solves the clustering task, and iii) integration of multicenter cohorts. These advances enable us to uncover five robust clinical states in pneumonia, three strongly associated with disease outcomes and two reflecting disease etiology.

Pneumonia is the world's leading cause of death, posing a significant burden to healthcare systems. In the United States, pneumonia is the second-most common cause of hospital admission, and around 20% of adult pneumonia hospitalizations include at least one intensive care unit (ICU) stay (1). For patients admitted to the ICU for other causes, pneumonia is the most common secondary complication, with an attributable mortality of around 10% (2). Although several pneumonia classification schemes have been proposed (3), risk factors identified (1, 4), and severity scores developed (5, 6), many challenges remain. Diagnosis and treatment prognosis for severe pneumonia remains difficult, ambiguous, and subject to antibiotic abuse (2, 7). This is largely due to the complex nature of the disease—pneumonia is intrinsically heterogeneous, characterized by a variety of pathologies, symptoms, comorbidities, and clinical courses. The myriad ways in which pneumonia evolves over time are poorly understood, with vastly different treatment responses under seemingly indistinguishable clinical manifestations (8). To identify meaningful phenotypic differences over pneumonia progression and understand patterns of states transition dynamics would shed light on precise clinical therapeutics and improve prognosis reliability.

Compared with traditional top–down, biomarker-based ways of characterizing pneumonia subtypes (4, 9, 10), data-driven methods hold the potential to capture the complexity of pneumonia trajectories in a novel way (11–13). Electronic health records (EHR) systems have been adopted worldwide by many healthcare providers (14). EHR contain rich information on patients' clinical phenotypes, including medical events, vital signs, medications, lab assays, diagnosis, and so on (15). Moreover, EHR track large populations of patients over long periods of time, thus serving as a unique instrument to uncover heterogeneity and progression trajectories of diseases (14, 16, 17). Recently, EHR-based studies have begun to shed light on biomarker identification (18), patient stratification (19, 20), risk prediction (21), and chronic disease management (16, 22). However, the potential of EHR comes with challenges, highlighted by irregularity, heterogeneity, and sparsity of the data (23, 24). Specifically, EHR contains multimodal

Author contributions: F.X. and L.A.N.A. designed research; F.X. performed research; F.X. contributed new reagents/analytic tools; F.X. and F.L.M. analyzed data; and F.X. and L.A.N.A. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

Copyright © 2024 the Author(s). Published by PNAS. This article is distributed under Creative Commons Attribution-NonCommercial-NoDerivatives License 4.0 (CC BY-NC-ND).

[1] To whom correspondence may be addressed. Email: amaral@northwestern.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2417688121/-/DCSupplemental.

Published October 30, 2024.

Downloaded from https://www.pnas.org by "NORTHWESTERN UNIVERSITY, SERIALS DEPARTMENT" on November 1, 2024 from IP address 165.124.84.65.

clinical measurements that are irregularly sampled with uneven time intervals, with high levels of missing entries that are not random but may reflect conscious decisions by clinicians. Despite efforts being made to accommodate the observational nature of EHR, most EHR-based studies still rely on heuristic strategies and ad hoc adjustments (15). A framework with objective criteria that standardizes EHR processing, including feature selection, normalization, missing value imputation, and dimensionality reduction, is not only necessary for robustly investigating pneumonia disease states but also helpful for the general field of EHR data mining (Fig. 1).

Our study aims to determine whether it is possible to identify in a data-driven manner clinical states associated with a pneumonia diagnosis from time discriminated EHR and, if the answer is positive, to characterize patterns of states transition dynamics (Fig. 1*B*). To this end, we develop and validate a semisupervised data preprocessing pipeline that optimizes feature selection, normalization, imputation, and dimensionality reduction toward the ability to distinguish patient-day pairs according to their associated mortality. We benchmark common clustering algorithms, improve clustering robustness via an ensemble strategy, and demonstrate the utility of ensemble density peak clustering (DPC) in both synthetic and real-world EHR datasets. Our novel, rigorous approach enable us to identify five clinical states, as well as verify their clinical significance. In addition, we are able to characterize patterns of state transition dynamics for different pneumonia etiologies and discharge dispositions.

Data and Methods

Data. The **SCRIPT** (25) dataset comprises the records of 585 patients with suspected severe pneumonia enrolled in Northwestern University Successful Clinical Response In Pneumonia Therapy (SCRIPT) study (https://script.northwestern.edu) between June 2018 and February 2022. The data are available in Physionet (26) under https://doi.org/10.13026/5phr-4r89. To be included in the cohort, a patient had 1) to be admitted to the ICU, 2) to have required mechanical ventilation, and 3) to have undergone at least one BAL, a procedure routinely conducted to support pneumonia diagnosis. We show in Fig. 1a a representative patient timeline from hospital admission to discharge.

The MIMIC-IV (Medical Information Mart for Intensive Care IV) dataset (27) is a comprehensive, general-purpose dataset that contains structured and unstructured EHR of more than 40,000 patients admitted to the ICU of Beth Israel Deaconess Medical Center between 2008 and 2019. To ensure consistency with the SCRIPT dataset, we restrict our analysis of MIMIC-IV data to the structured EHR records of a patient's ICU stay. We identify pneumonia patients according to the International Classification of Diseases (ICD) codes (28). In total, we identified 9,918 pneumonia patients from MIMIC-IV who required mechanical ventilation and were admitted to the ICU.

The patients in the SCRIPT and MIMIC-IV cohorts differ in a number of important ways (Table 1). First, due to its sampling time window, nearly one third of the SCRIPT cohort have a diagnosis of COVID-19 patients, while none of the patients in the MIMIC-IV cohort have such a diagnosis. Second, patients in the SCRIPT cohort are, on average, more severely ill than those in the MIMIC-IV cohort. This is visible in their significantly higher overall mortality rates, higher admission sequential organ failure assessment (SOFA) scores, and longer ICU stays. Third, the records of the patients in the SCRIPT cohort have more detailed annotations and adjudications than those of the MIMIC-IV cohort. For example, patients in the SCRIPT cohort more frequently have their pathogens determined (84.1% vs. 39.9%), and have manually adjudicated pneumonia episodes.

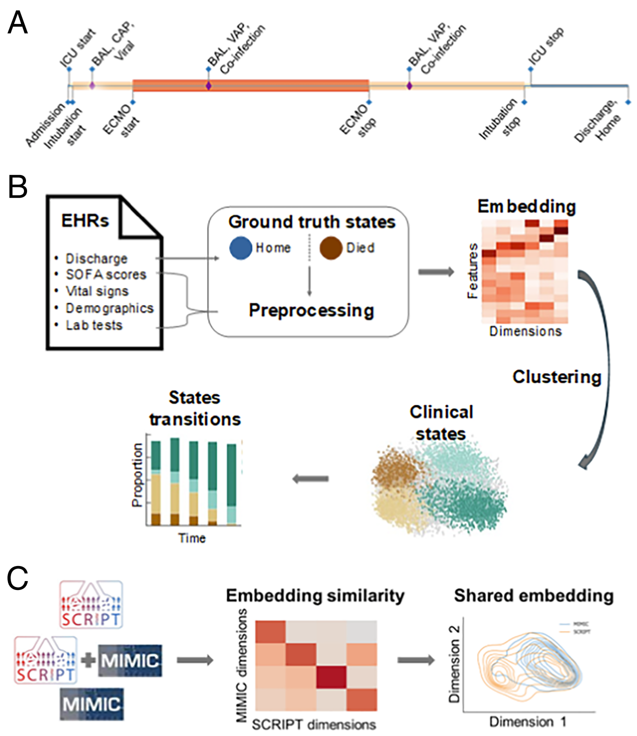

Fig. 1. Illustration of study workflow. (*A*) Hospitalization timeline for a representative patient with one ICU stay (thin gray line), who undergoes mechanical ventilation (light yellow bar) and extracorporeal membrane oxygenation (ECMO; light red bar). Before discharge to home, the patient spends some days in the ward (thicker gray bar). The patient undergoes three broncho-alveolar lavages (BALs, purple diamond), which yield diagnoses of community-acquired pneumonia (CAP) with viral infection (first BAL) and ventilator-associated pneumonia (VAP) with bacterial and viral coinfection (second and third BALs). (*B*) Illustration of data processing at single center level. We extract clinical features from structured EHRs. We then identify a common-sense ground-truth and select preprocessing steps that balance maximization of discrimination of extreme states while minimizing data loss. We learn a low-dimensionality embedding space using PCA and use the described ensemble DPC approach to reliably and robustly identify clinical states. We associate those clinical states with patient outcomes and with disease etiology by studying transitions between clinical states. (*C*) To integrate multicenter cohorts, we identify common features, characterize similarities of embedding spaces, and determine the embedding that provides the richest characterization of the data.

Data Preprocessing. The SCRIPT dataset presents structured EHR for each patient's patient-day pairs of their ICU stay (29). We group clinical features according to their nature and level of completeness (*SI Appendix*, Fig. S1).

Multimodal clinical features extracted from the MIMIC-IV dataset are irregular in nature, with various sampling frequencies and lengths. As described in ref. 29 with the SCRIPT dataset, we regularize all time series to a per-day basis. To facilitate research reproducibility, we use public scripts from the MIMIC Code Repository (30) to extract and curate clinical features whenever available.

For both datasets, we encode patient final discharge dispositions into 3 categories: "Dying" (patients who pass away), "Recovered" (sent home, excluding "Home for hospice"), and "Other." The latter includes discharges to other healthcare facilities such as long-term acute care hospitals (29).

Downloaded from https://www.pnas.org by "NORTHWESTERN UNIVERSITY, SERIALS DEPARTMENT" on November 1, 2024 from IP address 165.124.84.65.

Table 1. Summary of cohort patient characteristics

		SCRIPT		MIMIC-IV	
Number of patients		585		9,918	
Age, mean (SD)		60.6	(15.2)	66.6	(16.3)
Gender, n (%)	Male	346	(59.1)	5,649	(57.0)
	Female	239	(40.9)	4,269	43.0)
Ethnicity, n (%)	White	344	(58.8)	6,579	(66.3)
	Black or African American	119	(20.3)	1,086	(10.9)
	Asian	17	(2.9)	287	(2.9)
	Other			788	(7.9)
	Unknown or Not Reported	105	(17.9)	1,178	(11.9)
Discharge, n (%)	Died	243	(41.5)	2,112	(21.3)
	Other	209	(35.7)	5,021	(50.7)
	Home	133	(22.7)	2,771	(28.0)
Patient category, n (%)	COVID-19	190	(32.5)		
	Other Viral Pneumonia	50	(8.5)	429	(4.3)
	Nonviral Pneumonia	252	(43.1)	3,524	(35.5)
	Nonpneumonia control	93	(15.9)		
	Undetermined			5,965	(60.1)
Number of ICU stays, mean (SD)		1.3	(0.7)	1.2	(0.5)
Number of ICU days, mean (SD)		21.4	(22.9)	9.5	(9.9)
Admit SOFA score, mean (SD)		10.6	(4.2)	5.5	(3.7)

Ground Truth. We develop a semisupervised pipeline to optimize data preprocessing in a manner that is as unbiased, objective, and as automated as possible. We start by creating an approximate ground truth via filtering and partitioning patient-days into what most people would agree are vastly distinct extreme states—patient-days near patient death vs. patient-days near patient discharge to home. Specifically, we extract ICU patient-days within 48 h of patient death, as death typically occurs in the ICU, and ICU patient-days within 10 d of discharge to home, as the patient may be moved to ward prior to discharge. The former defines the ground truth Dying state while the latter defines the Recovered state. As a sanity test, we reason that an appropriate analysis pipeline should be at least able to distinguish these two extreme and vastly different states.

We quantify the separation of the two ground truth states by the silhouette coefficient (31) and quantify performance of a support vector machine (SVM) in predicting these extreme states from the learned embedding by area under the receiver operating characteristic (AUROC) curve. Thereby, we optimize the choices of feature selection, normalization, imputation, and dimensionality reduction in a semisupervised manner, toward better separation of extreme states that we can safely assume to be vastly different.

Feature Normalization. Clinical features in EHR are of vastly different scales and can be numerical (discrete or continuous), or categorical (nominal or ordinal) in nature. The challenge thus is to choose a feature normalization strategy that harmonizes such heterogeneity while preserving physiologically meaningful variance among patient-day vectors. To tackle this challenge, we benchmark untransformed (raw) data against five commonly used normalization strategies (*SI Appendix*, Fig. S2): MinMax scaler (MM), Standard Scaler (SS), Robust Scaler (RS), k-bin discretizer (KBD), and a combination of KBD and MinMax Scaler (KBD+MM).

For the SCRIPT cohort, as shown in *SI Appendix*, Fig. S2*A*, KBD continuously achieves high silhouette coefficients for prevalent feature sets, and only when including ventilation or lab test features with more sparsity does its performance drop significantly. Thus, we use KBD for feature normalization and consider feature sets up comprising combinations of the set of SOFA subscores, vital signs, demographics, and frequent lab tests in the downstream analysis of the SCRIPT cohort.

For the MIMIC cohort, as shown in *SI Appendix*, Fig. S2*B*, the MM scaler continuously achieves high Silhouette coefficients for various feature sets. The set of SOFA subscores is the only feature set where the MM scaler performs marginally worse than KBD. Thus, we use the MM scaler for the MIMIC cohort.

Missing Value Imputation. The nonrandom nature of missing values in the EHR requires suitable approaches to missing value imputation in order to avoid artifacts (32). We benchmark four commonly used data imputation methods: last observation carried forward (LOCF), k-nearest neighbors (KNN), multiple imputation by chained equations (MICE), and fill zero (FZ).

For the SCRIPT cohort, as shown in *SI Appendix*, Fig. S3*A*, LOCF with limit 2 improves data availability without significantly compromising the discrimination performance. On the other hand, for the MIMIC-IV cohort, all imputation strategies significantly deteriorate our ability to discriminate between the two extreme clinical state (*SI Appendix*, Fig. S3*B* for details). Therefore, we use LOCF with limit 2 to impute missing values in the SCRIPT dataset, while for the MIMIC dataset, we do no additional imputation of missing values.

Dimensionality Reduction and Feature Selection. We examine to what extent current machine learning algorithms can learn from sparse and noisy features. To this end, we examine the learned principal component analysis (PCA) (33) output as we gradually include more features. We choose PCA because it is a linear transformation that does not introduce uncontrolled nonlinear deformations of the data. This allows us to see and interpret what PCA learns from each set of features. We determine the number of significant principal components (PCs) using Horn's parallel analysis (34). We determine whether additional features are informative or noisy by evaluating the top feature loadings of significant PCs. Newly added features are deemed noninformative or noisy if adding them does not lead to significant change in the top feature loadings (*SI Appendix*, Figs. S5 and S7). We quantify similarity among

Downloaded from https://www.pnas.org by "NORTHWESTERN UNIVERSITY, SERIALS DEPARTMENT" on November 1, 2024 from IP address 165.124.84.65.

PCs derived from different feature sets by cosine similarity (*SI Appendix*, Fig. S6).

For the SCRIPT dataset, we identify four different feature sets that are informative (*SI Appendix*, Fig. S5), and four consensus PCs across feature sets as well as relatively distinctive ones with more features (*SI Appendix*, Fig. S6). For the MIMIC dataset, different feature sets result in very similar PCA spaces with only two significant PCs (*SI Appendix*, Fig. S8). These results indicate that for SCRIPT dataset, the four different feature sets may reveal slightly different heterogeneity of patient-day vectors. In contrast, the patients in MIMIC-IV dataset form a more homogeneous cohort, and the observed variability can be well explained solely by the SOFA subscores, additional features not significantly altering the learned embedding.

Clustering: Ensemble DPC (eDPC). Clustering of high-dimensional data is a remarkably challenging task. Many commonly used approaches are stochastic in nature and typically yield vastly different results for different executions of the same algorithm or different hyperparameter choices. Moreover, different clustering approaches frequently show very poor agreement in their outputs and produce clusters that do not conform to common sense expectations (such as a density peak at the cluster center; see *SI Appendix*, Figs. S11–S13).

DPC (35) was proposed as a way to identify cluster centers based on two common sense criteria. First, a cluster center must lie in a region of high local density. Second, a cluster center should not be too close to other cluster centers (to avoid overfitting). One can inspect these two criteria by calculating each data point's local density and its distance to the nearest point with a higher local density. Typically, one then identifies data points with high values for both criteria as density peaks and, thus, potential cluster centers.

While the DPC algorithm satisfies common sense expectations, there are several issues that arise when implementing it (36). First, local density estimation as originally implemented is quite noisy, and overly sensitive to hyperparameter choice. Second, while it is not known a priori how many peaks there are in the

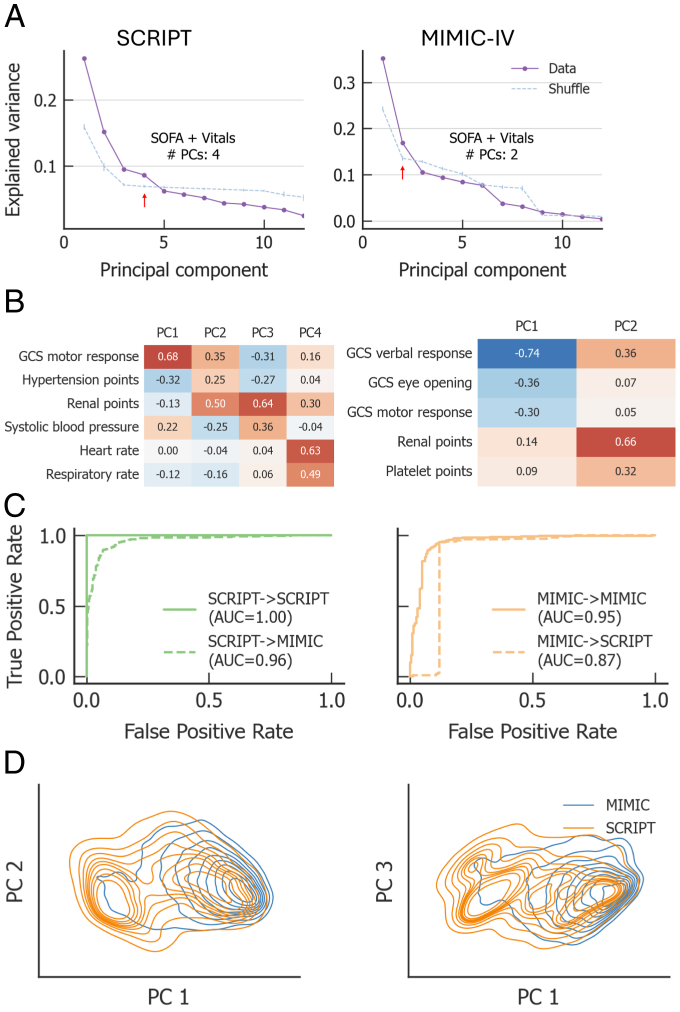

Fig. 2. Low-dimensionality embedding space learned from SOFA subscores and vital signs captures diversity of patient data. (*A*) We compare explained variance in data vs. their randomization to determine the number of significant PCs for SCRIPT (*Left*) and MIMIC-IV (*Right*) cohorts. We plot the fraction of variance explained by each PC for the data (purple solid line) and for shuffled data (blue dashed line). Error bars show 90% CI constructed by bootstrapping and the red arrow shows the last significant PC. (*B*) Top feature loadings of the significant PCs for SCRIPT and MIMIC cohorts. Red indicates positive loadings and blue indicates negative loadings. Greater color saturation indicates larger magnitude. (*C*) Within-distribution and out-of-distribution performance of models for discriminating extreme states. We compute the AUC for SVM models trained on the SCRIPT training dataset, using SOFA subscores and vital signs as features, on the SCRIPT test dataset (green solid line) and MIMIC-IV dataset (green dashed line). We find outstanding performance. We also compute the AUC for SVM models trained on the MIMIC-IV training dataset, using SOFA subscores and vital signs as features, on the MIMIC-IV test dataset (orange solid line) and on the SCRIPT dataset (orange dashed line). We find good but lower performance. (*D*) Projections of combined distributions of patient-day vectors onto learned embedding spaces learned from the SCRIPT training dataset. It is visually apparent that the two cohorts have different characteristics.

Downloaded from https://www.pnas.org by "NORTHWESTERN UNIVERSITY, SERIALS DEPARTMENT" on November 1, 2024 from IP address 165.124.84.65.

data, especially for noisy real-world datasets with complex density distributions or high dimensionality, DPC provides no objective way to correctly identify that number. Third, the way of assigning cluster identity for the rest of data points is overly simplistic and ignores uncertainties within potentially overlapping regions.

We build upon DPC (35) by specifically addressing these three major issues and, in doing so, obtain a significantly increased performance (*SI Appendix*, Table S2). First, we use a Gaussian kernel, instead of hard-coded cutoff, for density estimation, and determine optimal bandwidth parameter via fivefold cross validation. Second, we repeatedly compute the density and distance for bootstrapped samples, and construct an aggregate list of candidate cluster centers from multiple bootstrapped samples. We then select as candidate cluster centers the top-ranked centers according to the product of distance and density. Third, we conduct K-means clustering (37) on candidate cluster centers and determine the optimal number of clusters via maximization of the silhouette coefficient (31). Fourth, we assign cluster centers to the center of mass of each K-means identified cluster. Fifth, we fit a Gaussian mixture model (38) with fixed centers on density peaks, thereby determining cluster identity for each data point, as well as quantifying the uncertainty of each cluster assignment.

Results

Integration of Two Datasets in Shared PCA Embedding. Using SOFA-subscores plus vital signs, we identify four and two significant PCs for the SCRIPT and MIMIC dataset, respectively (Fig. 2*A*). As shown in Fig. 2*B*, the two largest PCs are characterized by GCS scores and renal points, respectively, for both the SCRIPT and MIMIC datasets. Vital signs dominate the loadings for the other two significant PCs obtained for the SCRIPT dataset.

To further validate that the data preprocessing pipeline captures the ground truth variance between extreme Dying and Recovered states, we train SVMs models using the learned PCA representations. We then test its performance in predicting the ground-truth Dying vs. Recovered states in both within-distribution and out-of-distribution test sets. Specifically, we train the SVM model on the SCRIPT training set (9:1 train-test splitting ratio) and test it on: 1) the SCRIPT test set (within-distribution test), 2) the MIMIC-IV dataset (out-of-distribution test). We also train the SVM model on the MIMIC-IV training set 5:5 train-test splitting ratio) and test it on: 1) the MIMIC-IV test set (within-distribution test), 2) the SCRIPT training set (out-of-distribution test).

We find that SCRIPT-trained SVM achieves an AUROC near 1 for both in-distribution and out-of-distribution test sets (Fig. 2*C* and *SI Appendix*, Fig. S9). In fact, the SVM model trained on SCRIPT performs as well as on MIMIC-IV training set as SVM model trained on it. In contrast, the SVM model trained on MIMIC-IV cannot perform as well on the SCRIPT dataset. This suggests that the PCA representation learned with SCRIPT captures the underlying variance and structure of MIMIC-IV data, but not vice versa. Therefore, we project the MIMIC-IV data onto the PCA space learned from the SCRIPT cohort (Fig. 2*D*).

eDPC Robustly Identifies Five Clinical States. Upon inspection of data distribution along each dimension, it is visually apparent that the distribution of patient-day pairs in PC space is nonuniform and that the two cohorts occupy both overlaying and distinct regions in the shared PCA space (Fig. 2*D* and *SI Appendix*, Fig. S10). This observation suggests the existence of discrete, high-density regions (clusters) that may correspond to physiologically distinct clinical states.

We use eDPC to reliably identify clusters that correspond to clinical states from patient-day state vectors in the reduced feature space (*SI Appendix*, Text). We identify consensus clustering solutions from the SCRIPT dataset with four different feature sets, although at different levels of granularity (*SI Appendix*, Figs. S16–S19). For the MIMIC dataset, eDPC consistently identifies only two clusters with different feature sets, while the GMM model fitted with SCRIPT data achieves as well, if not better, a mortality differentiation (*SI Appendix*, Fig. S20). We thus focus on the five-clusters solution obtained from SCRIPT training cohort with SOFA subscores plus vital signs features and examine its generalization to in-distribution (SCRIPT testing) and out-of-distribution (MIMIC) cohorts. Indeed, the GMM model fitted with SCRIPT training data identifies similar five clusters in the PCA space (i.e., similar feature characteristics) from both SCRIPT testing and MIMIC cohorts (Fig. 3*A*).

Uncovered States and Transitions Are Clinically Meaningful. To further validate the clinical relevance of the five uncovered clinical states, as well as investigate the dynamics of pneumonia progression, we break down each ICU stay into quintiles and separate patients according to disease outcome and etiology (Fig. 3*B*).

Across the SCRIPT training, testing, and MIMIC-IV cohorts, we observe an increasing proportion, along their ICU stay, of patient-day vectors classified into cluster C5 (golden brown) for patients who died, and a decreasing proportion of patient-day vectors classified into cluster C5 for patients who recovered

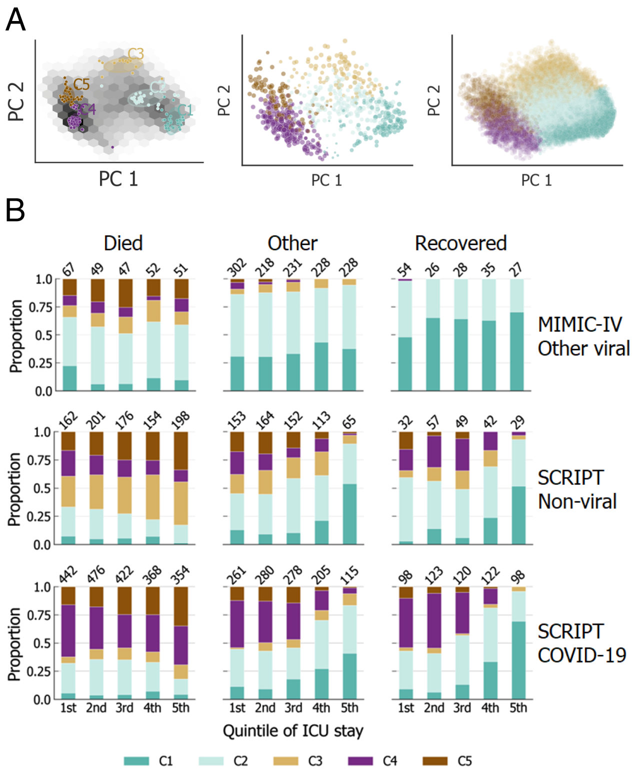

Fig. 3. Clinical states are associated with patient outcomes and disease etiology. (*A*) GMM model trained on SCRIPT training dataset (*Right* panel). Cluster membership of patient-day vectors for the SCRIPT testing dataset (*Middle* panel) and the MIMIC-IV cohort (*Left* panel). (*B*) Proportion of patient-day vectors classified into each of the five clinical states for patients stratified by patient outcome and disease etiology. Note strong association of clusters C1, C2, and C5 to patient outcomes and of cluster C4 to COVID-19.

Downloaded from https://www.pnas.org by "NORTHWESTERN UNIVERSITY, SERIALS DEPARTMENT" on November 1, 2024 from IP address 165.124.84.65.

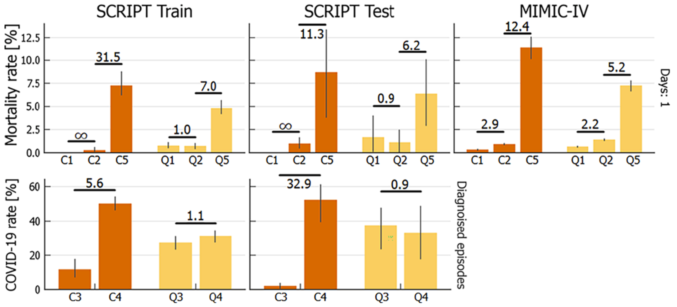

Fig. 4. Clinical states yield greater discriminatory power of patient outcomes than SOFA scores at short time horizons and provide earlier insight into disease etiology. (*Top* row) Next-day mortality rates for patients with patient-day vectors in clusters C1, C2, or C5 and or SOFA score in quintiles Q1, Q2, or Q5. Number over bars show ratio of mortality rates between two groups. Across all three datasets, clinical states provide greater stratification of patient mortality than SOFA scores. (*Bottom* row) Percentage of patients with a COVID-19 episode diagnosis (see *SI Appendix*, Text for details) for patients with a patient-day vector in clusters C3 or C4 and or SOFA score in quintiles Q3 or Q4. Number over bars show ratio of COVID-19 diagnosis rates between two groups. Patients diagnosed with COVID-19 are overrepresented on cluster C4.

Downloaded from https://www.pnas.org by "NORTHWESTERN UNIVERSITY, SERIALS DEPARTMENT" on November 1, 2024 from IP address 165.124.84.65.

(Fig. 3*B* and *SI Appendix*, Figs. S22–S25). In contrast, we find an increasing proportion, along their ICU stays, of patient-day vectors classified into cluster C1 (jade) for patients who did not die, and a decreasing proportion of patient-day vectors classified into clusters C1 and C2 (turquoise) for patients who died (see also *SI Appendix*, Figs. S18–S20).

While states C1, C2, and C5 are associated with disease outcome, states C4 and C3 are associated with etiology Fig. 3*B* and *SI Appendix*, Figs. S22–S25). Patients with a COVID-19 pneumonia diagnosis—but not other viral pneumonia—are characterized by patient-day vectors that are preferentially classified into cluster C4 (purple) and appear to be excluded from cluster C3 (gold). In contrast, patients with a nonviral pneumonia diagnosis—and for MIMIC-IV, especially patients with fungal infections—are characterized by patient-day vectors that are preferentially classified into cluster C3.

Mortality Rate Discrimination. An important question remaining to be answered is whether the uncovered clinical states provide greater discriminatory power with regard to mortality than the current gold standard—SOFA scores. To answer this question, we again stratify patients into groups with distinct SOFA scores, mortality rates, and etiologies. First, we look at mortality. Mortality rate is expected to increase linearly with SOFA score for long time horizons. However, at short time horizons, the SOFA score does not provide good discrimination (mortality rate of Q5 vs. Q2 is only about a factor of 5 to 7 and there is no discrimination between the mortality rates of Q2 and Q1). The mortality rate discrimination of the clinical states is considerably and significantly higher (Fig. 4 and *SI Appendix*, Fig. S30 and Table S4).

A second question remaining to be answered is whether the uncovered clinical states provide discriminatory power with regard to etiology. We can only use the SCRIPT training and testing cohorts for answering this question. For both, COVID-19 infection is consistently associated with cluster C4 but not with cluster C3. As a control, we can see that neither intermediate quintile of SOFA scores is associated with COVID-19 (Fig. 4 and *SI Appendix*, Fig. S31 and Table S5). Remarkably, the association with cluster C4 is indicative of COVID-19 etiology even before a formal diagnosis is registered, suggesting the potential for early prediction and patient stratification when the pathogen is unknown. This is especially significant given the large proportion of pneumonia patients whose pathogen remains unidentified throughout treatment.

Discussion

We identify and characterize five clinical states from structured EHR of pneumonia ICU patients in a robust data-driven manner. These five clinical states are associated with pneumonia etiology as well as disease outcomes. Our findings provide valuable insights for the integration of EHRs across different hospitals, especially when dealing with cohorts that differ on multiple aspects (Table 1). Although the MIMIC-IV cohort includes a significantly larger number of patients compared to the SCRIPT dataset, it exhibits less diversity especially on the more severe end of the spectrum. This lack of richness limits the granularity of the clinical states that can be identified from it alone. On the other hand, the diversity exhibited within the SCRIPT cohort enables models trained on SCRIPT data to achieve high discriminatory power and generalizability to the MIMIC-IV cohort. These findings underscore the importance of data sampling coverage, complexity, and quality over sheer sample size.

Our study has several limitations. First, our analysis only focuses on patient days spent within the ICU, thus not including mild pneumonia conditions or later recovery stages. Second, the five clinical states identified are unlikely to be exhaustive. Larger datasets and greater diversity of patients characteristics will likely reveal additional states. Third, we did not investigate how clinical interventions may affect state transitions. Fourth, we focused on aggregate characteristics and did not investigate individual trajectories. These are all important matters for future research.

Data, Materials, and Software Availability. The code supporting this study is available at Github (https://github.com/amarallab/Clinical_states) (39). Previously published data were used for this work (25–27, 40).

ACKNOWLEDGMENTS. Support by the National Heart Lung and Blood Institute (R01HL140362; L.A.N.A.), the National Institute of Allergies and Infectious Diseases (U19AI135964; L.A.N.A. and F.X.), and NIH Training Grant (T32GM153505; F.X.). Helpful discussions with T. Stoeger, C. Gao, R. Wunderink, and N. Markov.

Author affiliations: [a]Department of Engineering Sciences and Applied Math, Northwestern University, Evanston, IL 60208; [b]Interdisciplinary Biological Sciences Program, Northwestern University, Evanston, IL 60208; [c]Department of Medicine, Division of Pulmonary and Critical Care Medicine, Northwestern University School of Medicine, Chicago, IL 60611; [d]Department of Molecular Biosciences, Northwestern University, Evanston, IL 60208; [e]Department of Physics and Astronomy, Northwestern University, Evanston, IL 60208; [f]Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208; and [g]NSF-Simons National Institute on Theory and Mathematics in Biology, Northwestern University, Chicago, IL 60611

1. A. D. Storms *et al*., Rates and risk factors associated with hospitalization for pneumonia with ICU admission among adults. *BMC Pulm. Med.* **17**, 208 (2017).
2. A. C. Morris, Management of pneumonia in intensive care. *J. Emerg. Crit. Care Med.* **2**, 101 (2018).
3. G. Mackenzie, The definition and classification of pneumonia. *Pneumonia* **8**, 14 (2016).
4. J. F. Kronberger *et al*., Bronchoalveolar lavage and blood markers of infection in critically ill patients—A single center registry study. *J. Clin. Med.* **10**, 486 (2021).
5. G. Waterer, Severity scores and community-acquired pneumonia. Time to move forward. *Am. J. Respir. Crit. Care Med.* **196**, 1236–1238 (2017).
6. M. C. Vazquez Guillamet, M. H. Kollef, Next steps in pneumonia severity scores. *Clin. Infect. Dis.* **72**, 950–952 (2021).
7. I. S. Douglas, New diagnostic methods for pneumonia in the ICU. *Curr. Opin. Infect. Dis.* **29**, 197–204 (2016).
8. A. J. Quinton *et al*., Integrative physiology of pneumonia. *Physiol. Rev.* **98**, 1417–1464 (2018).
9. H. Zhou *et al*., Risk stratification and prediction value of procalcitonin and clinical severity scores for community-acquired pneumonia in ED. *Am. J. Emerg. Med.* **36**, 2155–2160 (2018).
10. S. H. Choi *et al*., Usefulness of cellular analysis of bronchoalveolar lavage fluid for predicting the etiology of pneumonia in critically ill patients. *PLoS ONE* **9**, e97346 (2014).
11. B. Gerlach *et al*., A robust data-driven approach identifies four personality types across four large data sets. *Nat. Hum. Behav.* **2**, 735–742 (2018).
12. T. Nagamine *et al*., Data-driven identification of heart failure disease states and progression pathways using electronic health records. *Sci. Rep.* **12**, 17871 (2022).
13. S. He, Z. Tian, A. Erdengasileng, N. Charness, "Temporal subtyping of Alzheimer's disease using medical conditions preceding Alzheimer's disease onset in electronic health records" in *AMIA Joint Summits of Translation Science Proceedings* (2022), p. 10.
14. P. B. Jensen, L. J. Jensen, S. Brunak, Mining electronic health records: Towards better research applications and clinical care. *Nat. Rev. Genet.* **13**, 395–405 (2012).
15. B. A. Goldstein, A. M. Navar, M. J. Pencina, J. P. A. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: A systematic review. *J. Am. Med. Inf. Assoc.* **24**, 198–208 (2017).
16. P. Yadav, M. Steinbach, V. Kumar, G. Simon, Mining electronic health records (EHRs): A survey. *ACM Comput. Surv.* **50**, 85 (2018).
17. J. C. Denny, Chapter 13: Mining electronic health records in the genomics era. *PLoS Comput. Biol.* **8**, e1002823 (2012).
18. A. Gupta, T. Liu, S. Shepherd, Clinical decision support system to assess the risk of sepsis using tree augmented Bayesian networks and electronic medical record data. *Health Inf. J.* **26**, 841–861 (2020).
19. I. Landi *et al*., Deep representation learning of electronic health records to unlock patient stratification at scale. *npj Digital Med.* **3**, 96 (2020).
20. D. Zeiberg *et al*., Machine learning for patient risk stratification for acute respiratory distress syndrome. *PLoS ONE* **14**, e0214465 (2019).
21. R. Caruana *et al*., "Intelligible models for HealthCare: Predicting pneumonia risk and hospital 30-day readmission" in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '15* (Association for Computing Machinery, 2015), pp. 1721–1730.
22. F. Xie *et al*., Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies. *J. Biomed. Inform.* **126**, 103980. (2022).
23. C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: A systematic review. *J. Am. Med. Inf. Assoc.* **25**, 1419–1428 (2018).
24. P. Y. Wu *et al*., -omic and electronic health record big data analytics for precision medicine. *IEEE Trans. Bio-Med. Eng.* **64**, 263–273 (2017).
25. N. Markov *et al*., SCRIPT CarpeDiem Dataset: demographics, outcomes, and per-day clinical parameters for critically ill patients with suspected pneumonia (version 1.1.0). PhysioNet (2023). https://doi.org/10.13026/5phr-4r89. Accessed 20 February 2024.
26. A. L. Goldberger *et al*., PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101**, e215–e220 (2000).
27. A. Johnson *et al*., MIMIC-IV (version 2.2). *PhysioNet* (2023). https://doi.org/10.13026/6mm1-ek67. Accessed 21 February 2024.
28. B. Wang *et al*., Novel pneumonia score based on a machine learning model for predicting mortality in pneumonia patients on admission to the intensive care unit. **217**, 107363 (2023).
29. C. A. Gao *et al*., Machine learning links unresolving secondary pneumonia to mortality in patients with severe pneumonia, including COVID-19. *J. Clin. Invest.* **133**, e170682 (2023).
30. A. E. W. Johnson, D. J. Stone, L. A. Celi, T. J. Pollard, The MIMIC code repository: Enabling reproducibility in critical care research. *J. Am. Med. Inf. Assoc.* **25**, 32–39 (2018).
31. P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
32. Y. Luo, Evaluating the state of the art in missing data imputation for clinical data. *Briefings Bioinf.* **23**, bbab489 (2022).
33. K. Pearson, LIII. On lines and planes of closest fit to systems of points in space. *London, Edinburgh, Dublin Philos. Mag. J. Sci.* **2**, 559–572 (1901).
34. J. L. Horn, A rationale and test for the number of factors in factor analysis. *Psychometrika* **30**, 179–185 (1965).
35. A. Rodriguez, A. Laio, Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014).
36. X. F. Wang, Y. Xu, Fast clustering using adaptive density peak detection. *Stat. Methods Med. Res.* **26**, 2800–2811 (2017).
37. S. Lloyd, Least squares quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).
38. D. Reynolds, "Gaussian mixture models" in *Encyclopedia of Biometrics*, S. Z. Li, A. Jain, Eds. (Springer US, 2009), pp. 659–663.
39. F. Xu, Clinical states. Github. https://github.com/amarallab/Clinical_states. Deposited 28 August 2024.
40. A. E. W. Johnson *et al*., MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).

Downloaded from https://www.pnas.org by "NORTHWESTERN UNIVERSITY, SERIALS DEPARTMENT" on November 1, 2024 from IP address 165.124.84.65.