Computational Endotypes of Sepsis and Septic Shock Derived from High Resolution Patient Records

Jason Guo BS[1,2], Han Kim MSE[1,3], Robert D Stevens MD[1,3]

[1]Johns Hopkins University Laboratory of Computational Intensive Care Medicine, [2]Department of Biology, Johns Hopkins University Krieger School of Arts and Sciences,
[3]Department of Anesthesiology and Critical Care Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA,

Introduction

Sepsis is a life-threatening syndrome characterized by organ dysfunction and dysregulated host response to infection.[1] Each year, sepsis accounts for up to 20% of deaths worldwide and $20 billion of healthcare costs in the US.[2,3] Sepsis and septic shock, a more severe subset, are identified using sepsis-3 consensus criteria[1], but the clinical and physiological variables of sepsis-3 may not capture the complexity of sepsis, which is characterized by a constellation of different clinical presentations, biological mechanisms, pathophysiological patterns, responses to treatment, and outcomes.

Rationale and Objectives

The intrinsic heterogeneity in sepsis could be the most significant barrier to developing effective therapies and predicting the onset, severity, time course, and outcomes for patients. By applying machine learning to high resolution patient records composed of granular physiologic time-series (PTS) data available in conjunction with electronic health records (EHRs), **we aim to identify novel endotypes of sepsis and septic shock.**

Methods

To identify adult patients in the Philips multi-center eICU database who met sepsis-3 criteria at any time, Sequential Organ Failure Assessment (SOFA) scores were calculated and infection state was determined through diagnosis and/or positive culture for each patient. Time-series features (sampling frequency every 5 min) were extracted. Patients were categorized every 5 minutes based on the Sepsis-3 criteria, in one of four categories: No Sepsis, Suspected Sepsis [only organ dysfunction], Sepsis, or Septic Shock during their ICU stay. We calculated 24 clinical and physiological features from 24 to 48 h after ICU admission. Features were scaled and used in unsupervised dimensionality reduction and clustering algorithms, accounting for mixed data types. Due to the large size of the time-series data, analysis was performed on a subsample of the full dataset, equalizing the number of samples in each state to 80% of the number of samples in the smallest state group. Groups identified by clustering (endotypes) were further characterized by ICU length of stay and hospital discharge outcomes.

Results

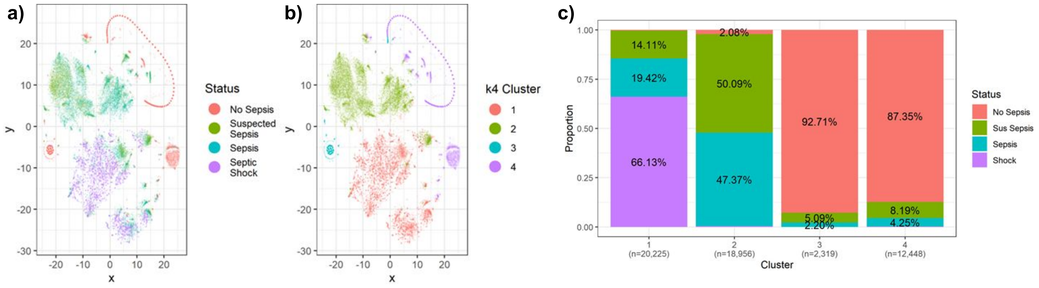

Figure 1. Identification and Characterization of Computational Endotypes. **a)** Visualization of patient cohorts, or "states," in scaled dataset containing 24 clinical and physiological variables using tSNE. Each point represents one hourly time point sample for each patient during the 24-48h interval after ICU admission. 13,487 timepoints (80% of septic shock timepoints) were randomly sampled for each state. **b)** Visualization of clusters from K-Means (k4) on scaled data overlapped on previous tSNE plot. **c)** Distribution of samples in each state calculated for every cluster. The number of samples in each cluster is given on the x-axis and only proportions larger than 0.5% are labelled.

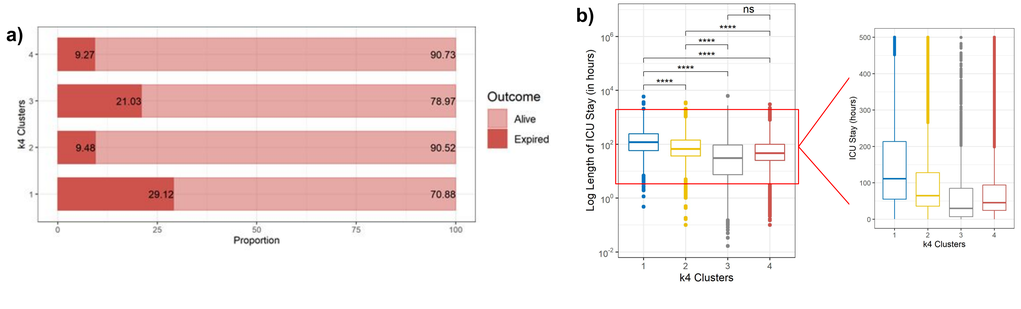

Figure 3. Outcome Distribution of Clusters. **a)** Distribution of patients surviving vs. deceased during hospital stay for each cluster. **b)** Log ICU length of stay (LOS) for each cluster (left panel). A portion of the raw data from the red outlined box is shown in the original linear scale (inset). Log transformation of LOS was done only for y-axis visualization and after hypothesis testing. $p \leq 0.0001$, ****; $p \leq 0.001$, ***; $p \leq 0.01$, **; $p \leq 0.05$, *; $p > 0.05$, ns.

Discussion

Unsupervised learning revealed four distinct clusters with distinct seps burdens and outcomes.

- Cluster 1 was the most severe cluster with the highest mortality and significantly longer LOS, containing all the "septic shock" patients
- Cluster 2 contained a high proportion of sepsis and suspected sepsis patients and had intermediate mortality
- Cluster 4 had a preponderance of "no sepsis" patients and the lowest mortality
- Clusters 3 and 4, although comparable in terms of sepsis/septic shock burdens, were differentiable in terms of mortality (21% vs 9% respectively)

Conclusions

Using only time-series data, we identified four candidate endotypes each associated with specific distributions of sepsis categories and specific outcome distribution. Ongoing research will further analyze identified endotypes in patient characteristics, outcomes, and response to specific treatments, as well as explore other machine learning algorithms and externally validate with other EHR data sources. In conjunction with a broader range of clinical and physiological features, these time-series-derived endotypes could help unlock the heterogeneity of sepsis, characterizing the evolution of its progression and yield greater precision in detection, prediction, and therapy. as a new paradigm for understanding and categorizing patients in the ICU who have sepsis.

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):801–810.
2. World Health Organization. Global report on the epidemiology and burden of sepsis: current evidence, identifying gaps, and future directions. https://apps.who.int/iris/bitstream/handle/10665/334216/9789240010789-eng.pdf. Published September 2020. Accessed September 17, 2020
3. Henry KE, Hager DN, Pronovost PJ, et al. A targeted real-time early warning score (TREWScore) for septic shock. *Sci Transl Med.* 2015;7(299): 299ra122..