Early Prediction of Sepsis from Clinical Data: the PhysioNet/Computing in Cardiology Challenge 2019

Matthew A Reyna[1], Chris Josef[1], Salman Seyedi[1], Russell Jeter[1], Supreeth P Shashikumar[2,3], M Brandon Westover[4], Ashish Sharma[1,*], Shamim Nemati[1,3,*], Gari D Clifford[1,2,*]

[1] Department of Biomedical Informatics, Emory University, USA
[2] Department of Biomedical Engineering, Georgia Institute of Technology, USA
[3] Department of Biomedical Informatics, University of California San Diego, USA
[4] Department of Neurology, Massachusetts General Hospital, USA
[*] The senior authors contributed equally to this manuscript.

Abstract

The PhysioNet/Computing in Cardiology Challenge focused on the early detection of sepsis from clinical data. A total of 40,336 patient records from two distinct hospital systems were shared with participants while 22,761 patient records from three distinct hospital systems were sequestered as hidden test sets. Each patient record contained up to 40 measurements of vital sign, laboratory, and demographics data for over 2.5 million hourly time windows and over 15 million data points. We used the Sepsis-3 clinical criteria to define the onset time of sepsis.

We challenged participants to design automated, open-source algorithms for predicting sepsis 6 hours before clinical recognition of sepsis. We developed a novel, clinical utility-based evaluation metric to assess each algorithm that rewards early sepsis predictions and penalizes late or missed predictions and false alarms.

A total of 104 teams from academia and industry submitted a total of 853 entries during the official phase of the Challenge. We accepted 90 abstracts based on Challenge entries for presentations at Computing in Cardiology. We also compared entries to ensure that approaches from different teams remained independent. This article presents our analysis and discusses the implications of the Challenge for early sepsis predictions and related sequential prediction tasks.

1. Introduction

The PhysioNet/Computing in Cardiology Challenge is an international competition for open-source solutions to complex physiological signal processing and medical classification problems [1]. In 2019, the Challenge's 20th year, we asked participants to develop automated techniques for the early detection of sepsis from clinical data [2].

Sepsis is a life-threatening condition that occurs when the body's response to infection causes tissue damage, organ failure, or death [3–5]. Nearly 1.7 million people develop sepsis and 270,000 die from sepsis each year in the U.S., and an estimated 30 million people develop sepsis and 6 million people die from sepsis each year globally [6]. In the U.S., managing sepsis is more expensive than any other health condition, where expenses exceed $24 billion annually or 13% of costs. Altogether, preventing and treating sepsis is a major public health issue with considerable morbidity, mortality, and healthcare costs [7–10].

Early detection and intervention are critical for improving outcomes of septic patients; each hour of delayed treatment is associated with 4-8% higher mortality [11, 12]. However, despite the introduction of new clinical criteria for recognizing sepsis [3–5], the fundamental need for early detection and treatment of sepsis remains unmet [13].

Computational approaches promise to improve early sepsis detection. Such approaches typically apply machine learning techniques to clinical data to make real-time predictions up to a day before clinical recognition of sepsis; see [14–16] for examples. However, these algorithms frequently address subtly different problems and are developed and tested in different patient cohorts with labels determined using different clinical criteria. We provided a common problem using multiple datasets and the same criteria for sepsis. Moreover, adequately describing such algorithms is a difficult task in a standard journal article. We encouraged participants to release their code in reproducible containers under open-source licenses. Finally, different studies often use varied evaluation metrics, and these metrics typically do not reward algorithms that facilitate early sepsis detection and treatment. We designed a novel metric that addresses this issue and could be generally applicable to infrequent events in sequential prediction tasks. Therefore, while computational approaches

demonstrate a potential for early sepsis predictions, the limits of such approaches remain unknown. The PhysioNet/Computing in Cardiology Challenge 2019 provides an opportunity to address these issues.

To discourage teams from designing algorithms with limited applicability, we imposed three key constraints on participants. First, we posted data from two separate hospital systems and sequestered data from a third system so that algorithms that overfit on the shared databases would underperform on the third database. Second, each team's algorithm was scored only once on the third database, preventing sequential training on the hidden data. Third, we evaluated the similarity between algorithms to identify teams that attempted to circumvent the rules by repeated scoring.

2. Challenge Data

2.1. Data Source

We sourced data for the Challenge from three geographically distinct U.S. hospital systems with three different electronic medical record (EMR) systems. These data were collected over the past decade with approval from the appropriate Institutional Review Boards. We deidentified and posted the data and labels for 40,336 patients from two of the three hospital systems as public training sets and sequestered the data and labels for 22,761 patients from three hospital systems as hidden test sets.

The Challenge data consist of 40 clinical variables, including 8 vital sign summaries, 26 laboratory values, and 6 patient descriptions; [2] provides a detailed discussion of the data.

2.2. Expert Labeling

We labeled the data using the Sepsis-3 clinical criteria [3–5]. For each septic patient, we identified three time points:

- $t_{\text{suspicion}}$: Clinical suspicion of infection is the earlier of intravenous (IV) antibiotics and blood cultures within a specified duration. If IV antibiotics were given first, then the cultures must have been obtained within 24 hours. If cultures were obtained first, then IV antibiotic must have been ordered within 72 hours. IV antibiotics must have been administered for at least 72 consecutive hours.
- t_{SOFA}: Occurrence of organ failure as identified by a two-point increase in the Sequential Organ Failure Assessment (SOFA) score within a 24-hour period.
- t_{sepsis}: Onset of sepsis is the earlier of $t_{\text{suspicion}}$ and t_{SOFA} as long as t_{SOFA} occurs no more than 24 hours before or 12 hours after $t_{\text{suspicion}}$.

Septic patients have $t_{\text{sepsis}} < \infty$, and non-septic patients have $t_{\text{sepsis}} = \infty$.

3. Challenge Objective

The goal of this Challenge is the design of algorithms for early predictions of sepsis using routinely available clinical data. We asked participants to design and implement working, open-source algorithms that can, based only on the provided clinical data, automatically identify a patient's risk of sepsis and make a positive or negative prediction of sepsis for every hourly time window in the patient's clinical record. In particular, we asked participants to predict sepsis at least 6 hours but no more than 12 hours before the onset of sepsis according to Sepsis-3 clinical criteria. To evaluate each algorithm, we designed a clinical utility-based scoring metric that prioritizes algorithms that make actionable predictions. The winners of the Challenge are the team whose algorithm gives predictions with the highest utility score on a hidden test set containing patient records from three hospital systems.

4. Challenge Scoring

We evaluated each algorithm's predictions using a novel metric that we created for this Challenge. To better capture the clinical utility of sepsis detection and treatment, this metric rewards algorithms for early sepsis predictions in septic patients, and it penalizes algorithms for late or missed sepsis predictions in septic patients and for sepsis predictions in non-septic patients.

Each algorithm makes a binary sepsis prediction for each hourly time window of each patient's record. We define a score for each prediction and aggregate these scores over all hourly time windows in all patient records to provide a score for the algorithm on a dataset.

Let $x(s,t) = 1$ indicate a positive sepsis prediction for patient s at time t and $x(s,t) = 0$ otherwise, and let $\delta(s) = 1$ if patient s is eventually septic and $\delta(s) = 0$ otherwise. We define a utility score

$$U(s,t) = \begin{cases} U_{\text{TP}}(s,t), & x(s,t) = 1,\ \delta(s) = 1, \\ U_{\text{FP}}(s,t), & x(s,t) = 1,\ \delta(s) = 0, \\ U_{\text{FN}}(s,t), & x(s,t) = 0,\ \delta(s) = 1, \\ U_{\text{TN}}(s,t), & x(s,t) = 0,\ \delta(s) = 0, \end{cases} \tag{1}$$

where $U_{\text{TP}}(s,t)$, $U_{\text{FP}}(s,t)$, $U_{\text{FN}}(s,t)$, and $U_{\text{TN}}(s,t)$ are illustrated in Fig. 1 for an example septic patient with sepsis onset $t_{\text{sepsis}} = 48$ and an example non-septic patient; the times in the plot are given as examples.

We compute a total utility score

$$U_{\text{total}} = \sum_{s \in S} \sum_{t \in T(s)} U(s,t) \tag{2}$$

for each algorithm and normalize it to define a normalized

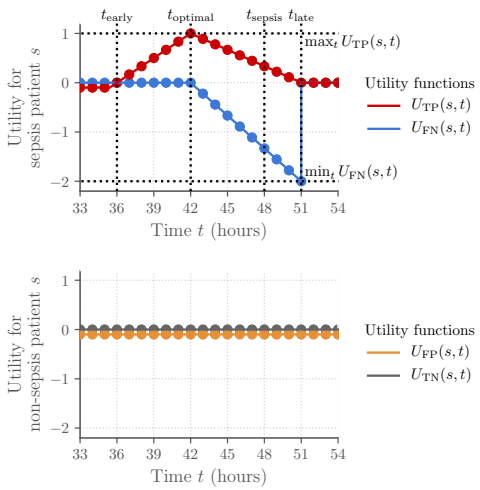

Figure 1. Utility of positive and negative sepsis predictions for septic (top) and non-septic (bottom) patients at each hourly time window in a patient's record; sepsis onset $t_{\text{sepsis}} = 48$ is given for an example septic patient.

utility score

$$U_{\text{normalized}} = \frac{U_{\text{total}} - U_{\text{no predictions}}}{U_{\text{optimal}} - U_{\text{no predictions}}} \tag{3}$$

so that an optimal algorithm achieving the highest possible score receives a normalized score of 1 and a completely inactive algorithm that makes only negative predictions receives a normalized score of 0.

Each team was allowed five scored submissions during an unofficial phase of the Challenge from 8 February 2019 to 19 April 2019 and ten scored submissions during an official phase from 25 April 2019 to 25 August 2019. Each algorithm received a normalized utility score (3) on the test set from hospital system A. The algorithm with the highest normalized utility score on the full test set from all three hospital systems wins.

4.1. Challenge Scoring Mechanism

Participants were required to submit their sepsis prediction algorithms through a cloud-based submission system. This approach encouraged reproducibility and gave participants the ability to validate their algorithms on real-world data without releasing the sequestered test data.

The submission system used containers that were orchestrated, as pipelines, on Google Cloud. Participants packaged their entries in Docker containers, and the submission system created a pipeline with the entry and our scoring function that it launched on Google Cloud.

Each entry was run in a virtual machine with 2 CPUs and 12 GB of RAM, and each entry was allowed 24 hours of run time on each hidden test set.

5. Results

A total of 104 teams from academia and industry submitted a total of 853 entries during the official phase of the Challenge with 430 successful entries from 88 teams. Each successful entry received its score on the test data for hospital system A, and each team nominated its favorite successful entry for evaluation on the test data for hospital systems B and C. Table 1 summarizes the teams with the highest-scoring entries. Unsurprisingly, algorithms generally performed worse on test data from hospital system C than hospital systems A and B.

6. Discussion

The PhysioNet/Computing in Cardiology Challenge 2019 asked participants to develop automated, open-source algorithms for the early detection of sepsis from clinical data. We assembled 63,087 patient records from three hospital systems. By posting data from two hospitals publicly and sequestering data from all three hospitals, we provided participants the opportunity to create training methodologies that do not overfit to one medical center. The third hidden database provided a strong indication of how well participants had accomplished this critical task.

We proposed and used a novel evaluation metric that captures the clinical utility of early sepsis detection, weighted by the relative "earliness" or "lateness" of the prediction. This metric could be considered for wider adoption in clinical care because it does not suffer from many of the problems of current metrics that either assume a one-shot decision (accuracy, F-measure, etc.) or no decision threshold (area under the curve metrics).

These efforts provide a more complete picture of how algorithms can provide early sepsis predictions.

Acknowledgments

This work was funded by the Gordon and Betty Moore Foundation. GC, SN, MR and RJ are partially funded by the National Science Foundation under award number 1822378 (Leveraging Heterogeneous Data Across International Borders in a Privacy Preserving Manner for Clinical Deep Learning). SN and SPS are also funded by the NIH award #K01ES025445. CJ is funded by the Surgical Critical Care Initiative (SC2i), sponsored by the Department of Defense Health Program Joint Program Committee 6 / Combat Casualty Care (USUHS HT9404-13-1-0032 and HU0001-15-2-0001). AS and the development of the cloud-based scoring system were partially supported by the National Cancer Institute (U24CA215109). This work was also supported by the National Center for Advancing Translational Sciences of the National Institutes of Health under Award Number UL1TR002378 and the Na-

Rank	Team	Final Score	Score A	Score B	Score C
1	James Morrill, Andrey Kormilitzin, Alejo Nevado-Holgado, Sumanth Swaminathan, Sam Howison, Terry Lyons	0.360	0.433	0.434	-0.123
2	John Anda Du, Nadi Sadr, Philip de Chazal	0.345	0.409	0.396	-0.042
3	Morteza Zabihi, Serkan Kiranyaz, Moncef Gabbouj	0.339	0.422	0.395	-0.146
4	Xiang Li, Yanni Kang, Xiaoyu Jia, Junmei Wang, Guotong Xie	0.337	0.420	0.401	-0.156
5	Janmajay Singh, Kentaro Oshiro, Raghava Krishnan, Masahiro Sato, Tomoko Ohkuma, Noriji Kato	0.337	0.401	0.407	-0.094
*	Meicheng Yang, Hongxiang Gao, Xingyao Wang, Yuwen Li, Jianqing Li, Chengyu Liu	0.364	0.430	0.422	-0.048

Table 1. Clinical utility scores for the teams with the five highest scores on the full test set from hospital systems A, B, and C (Final Score) as well as their scores on the separate test sets from hospital systems A, B, and C (Score A, Score B, and Score C, respectively). * denotes the highest-scoring unofficial entry.

tional Institutes of Health-sponsored Research Resource for Complex Physiologic Signals (www.physionet.org) (R01GM104987). The content of this article is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors declare no conflict of interest.

References

[1] Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 2000;101(23):e215–e220.

[2] Reyna MA, Josef C, Jeter R, Shashikumar SP, Westover MB, Nemati S, Clifford GD, Sharma A. Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology Challenge 2019. Critical Care Medicine 2019; In press.

[3] Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, et al. Assessment of clinical criteria for sepsis: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). Journal of the American Medical Association 2016;315(8):762–774.

[4] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). Journal of the American Medical Association 2016;315(8):801–810.

[5] Shankar-Hari M, Phillips GS, Levy ML, Seymour CW, Liu VX, Deutschman CS, Angus DC, Rubenfeld GD, Singer M. Developing a new definition and assessing new clinical criteria for septic shock: for the Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). Journal of the American Medical Association 2016;315(8):775–787.

[6] Organization WH. Sepsis. https://www.who.int/news-room/fact-sheets/detail/sepsis, 19 April 2018. [Online; accessed 1 February 2019].

[7] Angus DC, Linde-Zwirble WT, Lidicker J, Clermont G, Carcillo J, Pinsky MR. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. Critical Care Medicine 2001; 29(7):1303–1310.

[8] Martin GS, Mannino DM, Eaton S, Moss M. The epidemiology of sepsis in the United States from 1979 through 2000. New England Journal of Medicine 2003; 348(16):1546–1554.

[9] Moran J, Myburgh J, Syres G, Jones D, Cameron P, Higgins A, Finfer S, Webb S, Delaney A, Cross A, et al. The outcome of patients with sepsis and septic shock presenting to emergency departments in Australia and New Zealand. Critical Care and Resuscitation 2007;9(1):8.

[10] Stoller J, Halpin L, Weis M, Aplin B, Qu W, Georgescu C, Nazzal M. Epidemiology of severe sepsis: 2008-2012. Journal of Critical Care 2016;31(1):58–62.

[11] Kumar A, Roberts D, Wood KE, Light B, Parrillo JE, Sharma S, Suppes R, Feinstein D, Zanotti S, Taiberg L, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. Critical Care Medicine 2006; 34(6):1589–1596.

[12] Seymour CW, Gesten F, Prescott HC, Friedrich ME, Iwashyna TJ, Phillips GS, Lemeshow S, Osborn T, Terry KM, Levy MM. Time to treatment and mortality during mandated emergency care for sepsis. New England Journal of Medicine 2017;376(23):2235–2244.

[13] Prescott HC, Iwashyna TJ. Improving Sepsis Treatment by Embracing Diagnostic Uncertainty. Annals of the American Thoracic Society 2019;16(4):426–429.

[14] Henry KE, Hager DN, Pronovost PJ, Saria S. A targeted real-time early warning score (TREWScore) for septic shock. Science Translational Medicine 2015; 7(299):299ra122–299ra122.

[15] Nemati S, Holder A, Razmi F, Stanley MD, Clifford GD, Buchman TG. An interpretable machine learning model for accurate prediction of sepsis in the ICU. Critical Care Medicine 2018;46(4):547–553.

[16] Cheng LF, Prasad N, Engelhardt BE. An optimal policy for patient laboratory tests in intensive care units. In Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing, volume 24. World Scientific, 2019; 320–331.

Address for correspondence:

Matthew Reyna. DBMI, 101 Woodruff Circle, 4th Floor East, Atlanta, GA 30322. matthew.a.reyna@emory.edu