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

Matthew A. Reyna, PhD[1], Christopher S. Josef, MD[1], Russell Jeter, PhD[1], Supreeth P. Shashikumar[2,3], M. Brandon Westover, MD, PhD[4], Shamim Nemati, PhD[1,3,*], Gari D. Clifford, DPhil[1,2,*], and Ashish Sharma, PhD[1,*]

[1]Department of Biomedical Informatics, Emory University

[2]Department of Biomedical Engineering, Georgia Institute of Technology

[3]Department of Biomedical Informatics, University of California San Diego

[4]Department of Neurology, Massachusetts General Hospital

[*]The senior authors contributed equally to this manuscript.

14 October 2019

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. CSJ 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) and was supported by a grant from the National Institutes of Health, USA (NIH Grant 5T32GM095442-09). 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 National 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.

For information regarding this article, please contact MR. Address: Department of Biomedical Informatics, Emory University, Woodruff Memorial Research Building, 101 Woodruff Circle, 4th Floor East, Atlanta, GA 30322, USA. Phone: (404)-727-4631. Email: matthew.a.reyna@emory.edu.

Abstract

Objectives: Sepsis is a major public health concern with significant morbidity, mortality, and healthcare expenses. Early detection and antibiotic treatment of sepsis improve outcomes. However, while professional critical care societies have proposed new clinical criteria that aid sepsis recognition, the fundamental need for early detection and treatment remains unmet. In response, researchers have proposed algorithms for early sepsis detection, but directly comparing such methods has not been possible because of different patient cohorts, clinical variables and sepsis criteria, prediction tasks, evaluation metrics, and other differences. To address these issues, the PhysioNet/Computing in Cardiology Challenge 2019 facilitated the development of automated, open-source algorithms for the early detection of sepsis from clinical data.

Design: Participants submitted containerized algorithms to a cloud-based testing environment, where we graded entries for their binary classification performance using a novel clinical utility-based evaluation metric. We designed this scoring function specifically for the Challenge to reward algorithms for early predictions and penalize them for late or missed predictions and for false alarms.

Setting: ICUs in three separate hospital systems. We shared data from two systems publicly and sequestered data from all three systems for scoring.

Patients: We sourced over 60,000 ICU patients with up to 40 clinical variables for each hour of a patient's ICU stay. We applied Sepsis-3 clinical criteria for sepsis onset.

Measurements and Main Results: 104 groups from academia and industry participated, contributing 853 submissions. Moreover, 90 abstracts based on Challenge entries were accepted for presentation at Computing in Cardiology.

Conclusions: Diverse computational approaches predict the onset of sepsis several hours before clinical recognition, but generalizability to different hospital systems remains a challenge.

Key Words: sepsis; early detection and treatment; generalizability; evaluation metrics; sequential prediction tasks; open-source algorithms; competition; PhysioNet

1 Introduction

Sepsis is a life-threatening condition that occurs when the body's response to infection causes tissue damage, organ failure, or death [1, 2, 3]. In the U.S., nearly 1.7 million people develop sepsis and 270,000 people die from sepsis each year; over one-third of people who die in U.S. hospitals have sepsis [4]. Globally, an estimated 30 million people develop sepsis and 6 million people die from sepsis each year [5]. Costs for managing sepsis in U.S. hospitals exceed those for any other health condition at $24 billion annually (13% of U.S. healthcare expenses); a majority of these costs are for patients who develop sepsis during their hospital stay [6]. The developing world faces additional expenses from sepsis management and higher risks of adverse outcomes. Altogether, sepsis is a major public health issue responsible for significant morbidity, mortality, and healthcare expenses [7, 8, 9, 10].

The reliable and early identification of sepsis is often complicated by its syndromic nature, which can contribute to delays in treatment. The importance of early identification and treatment of sepsis is highlighted in two recent studies that suggest an increase in the adjusted mortality of septic patients who experienced delays in antibiotic therapy [11, 12]. This effect is even more profound in patients suffering from septic shock, where hourly delays were associated with an 3.6-9.9% increase in mortality per hour [13]. Professional critical care societies have proposed clinical criteria for recognizing and treating sepsis [1, 2, 3]; however, the fundamental need for early and reliable identification of sepsis remains unmet [14].

The PhysioNet/Computing in Cardiology Challenge is an international competition focused on open-source solutions for complex physiological signal processing and medical classification problems [15]. In 2019, the Challenge's 20th year, we asked participants to develop automated techniques for the early detection of sepsis from clinical data.

Computational approaches promise to improve the early detection of sepsis. Such approaches typically apply machine learning techniques to clinical data, e.g., [16, 17, 18], with

the goal of making real-time predictions up to a day before clinical recognition of sepsis. However, the relative strengths and weaknesses of algorithmic approaches are unclear for a variety of reasons. The PhysioNet/Computing in Cardiology Challenge 2019 provided an opportunity to explore the limits of such approaches.

First, algorithms for the early detection of sepsis frequently address subtly different problems, and they tend to have been developed and tested in different patient cohorts with different clinical variables and labels arising from different clinical criteria for sepsis. For the Challenge, we provided a common problem statement using the same clinical variables and sepsis criteria. We shared data from two separate hospital systems and sequestered data from a third hospital system. Algorithms that overfit on the shared databases typically underperformed on the hidden database, particularly if they encoded data collection behaviors specific to a given hospital system. Moreover, we ran algorithms only once on the full hidden dataset to prevent sequential training on the hidden data, and we compared algorithms to identify teams that attempted to circumvent the rules and have more "bites of cherry" than other teams.

Second, different studies often employ different evaluation metrics, and such metrics do not necessarily reflect the clinical utility of sepsis detection and treatment. Traditional scoring metrics, such as area under the curve (AUC) metrics, do not explicitly reward early detection or penalize false alarms or overtreatment. For the Challenge, we devised a novel evaluation metric that addresses these issues and could be generally applicable to predicting infrequent events in time series data.

Third, the complexity of such algorithms is nearly impossible to adequate describe in a research article. For the Challenge, we encouraged and facilitated the open sourcing of algorithms to ensure that subtle implementation details are provided and reproducible.

In this paper, we begin with the Challenge objective of early predictions of sepsis, the Challenge data and clinical criteria for sepsis, and a new evaluation metric that reflects

the clinical utility of early sepsis predictions. We continue with the Challenge submission procedure, the results of the Challenge, and a discussion of computational approaches for early predictions of sepsis.

2 Methods

2.1 Challenge Objective

The goal of this Challenge was the development of algorithms for the early prediction of sepsis using routinely available clinical data. Early predictions of sepsis are potentially life saving, while late or missed predictions are potentially life threatening, and false alarms consume hospital resources and erode trust in the algorithms themselves [19].

For this Challenge, 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 six hours (but no more than twelve hours) before the onset time of sepsis according to Sepsis-3 clinical criteria [1, 2, 3]. To evaluate each algorithm, we designed a new clinical utility-based scoring metric that rewards algorithms for early sepsis predictions and penalizes them for late and missed sepsis predictions as well as for false alarms. The winners of this Challenge were the team whose algorithm gave predictions with the highest clinical utility score for patients in a hidden test set across three hospital systems.

We awarded prizes to teams with winning algorithms. While we allowed both non-commercial and commercial entities to enter, only open-source entries were eligible for prizes. All code was required to be submitted to ensure that methods were replicable and because no teams had access to the hidden data. This allowed for the comparison of winning teams with commercial entities and increased the competitive landscape.

2.2 Challenge Data

We obtained the data for the Challenge from three geographically distinct U.S. hospital systems with three different electronic medical record systems: Beth Israel Deaconess Medical Center (hospital system A), Emory University Hospital (hospital system B), and a third, unidentified hospital system (hospital system C). These data were collected over the past decade with approval from the appropriate Institutional Review Boards. We de-identified and labeled the data using Sepsis-3 clinical criteria [1, 2, 3]. Data and labels for 40,336 patients from hospital systems A and B were posted publicly for download and data and labels for 24,819 patients from hospital systems A, B, and C were sequestered as hidden test sets.

The Challenge data consisted of a combination of hourly vital sign summaries, lab values, and static patient descriptions. In particular, the data contained 40 clinical variables: 8 vital sign variables, 26 laboratory variables, and 6 demographic variables; Table 1 describes these variables. Altogether, these data included over 2.5 million hourly time windows and 15 million data points.

Data extracted from the Electronic Medical Record (EMR) underwent a series of pre-processing steps prior to formal analysis and model development. All patient features were condensed into hourly bins simplifying model development and testing, e.g., multiple heart rate measurements in an hourly time window were summarized as the median heart rate measurement. Multiple Logical Observation Identifiers Names and Codes (LOINC) codes describing the same clinical parameter were condensed into a single variable, e.g., serum hemoglobin and arterial hemoglobin became hemoglobin.

We labeled patient data in accordance with Sepsis-3 clinical criteria [1, 2, 3]. For each septic patient, we specified the following three time points to define the onset time t_{sepsis} of sepsis:

	Measurement	Description
1	HR	Heart rate (beats per minute)
2	O2Sat	Pulse oximetry (%)
3	Temp	Temperature (deg C)
4	SBP	Systolic BP (mm Hg)
5	MAP	Mean arterial pressure (mm Hg)
6	DBP	Diastolic BP (mm Hg)
7	Resp	Respiration rate (breaths per minute)
8	EtCO2	End tidal carbon dioxide (mm Hg)
9	BaseExcess	Excess bicarbonate (mmol/L)
10	HCO3	Bicarbonate (mmol/L)
11	FiO2	Fraction of inspired oxygen (%)
12	pH	pH
13	PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
14	SaO2	Oxygen saturation from arterial blood (%)
15	AST	Aspartate transaminase (IU/L)
16	BUN	Blood urea nitrogen (mg/dL)
17	Alkalinephos	Alkaline phosphatase (IU/L)
18	Calcium	Calcium (mg/dL)
19	Chloride	Chloride (mmol/L)
20	Creatinine Creatinine	(mg/dL)
21	Bilirubin_direct	Direct bilirubin (mg/dL)
22	Glucose	Serum glucose (mg/dL)
23	Lactate	Lactic acid (mg/dL)
24	Magnesium	Magnesium (mmol/dL)
25	Phosphate	Phosphate (mg/dL)
26	Potassium	Potassiam (mmol/L)
27	Bilirubin_total	Total bilirubin (mg/dL)
28	TroponinI	Troponin I (ng/mL)
29	Hct	Hematocrit (%)
30	Hgb	Hemoglobin (g/dL)
31	PTT	Partial thromboplastin time (seconds)
32	WBC	Leukocyte count (count/L)
33	Fibrinogen	Fibrinogen concentration (mg/dL)
34	Platelets	Platelet count (count/mL)
35	Age	Age (years)
36	Gender	Female (0) or male (1)
37	Unit1	Administrative identifier for ICU unit (MICU); false (0) or true (1)
38	Unit2	Administrative identifier for ICU unit (SICU); false (0) or true (1)
39	HospAdmTime	Time between hospital and ICU admission (hours since ICU admission)
40	ICULOS	ICU length of stay (hours since ICU admission)
41	SepsisLabel	For septic patients, SepsisLabel is 1 if $t \geq t_{\text{sepsis}} - 6$ and 0 if $t < t_{\text{sepsis}} - 6$. For non-septic patients, SepsisLabel is 0.

Table 1: Clinical time series data used for the Challenge: vital signs (rows 1-8), laboratory values (rows 9-34), demographics (rows 35-40), and outcome (row 41).

- $t_{\text{suspicion}}$: Clinical suspicion of infection identified as the earlier timestamp of intravenous (IV) antibiotics and blood cultures within a given time interval. 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. In either case, 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 identified as the earlier of $t_{\text{suspicion}}$ and t_{SOFA} as long as t_{SOFA} occurred no more than 24 hours before or 12 hours after $t_{\text{suspicion}}$.

Missing and erroneous data were intentionally preserved as part of the Challenge. However, patients with less than 8 hourly time windows of data in the ICU were not included, and patients with t_{sepsis} less than 4 hours after ICU admission were not included. Patient records were truncated after ICU discharge, and patients with more than 2 weeks of hourly time windows were truncated to 2 weeks.

Table 2 summarizes the datasets for the two shared hospital databases. Fig. 1 shows the densities of entries (i.e., the fraction of non-empty hourly measurements) for each vital sign and laboratory value in each patient record; most vital signs were updated on an hourly basis in most patient records, and most laboratory values were updated on a daily basis. Fig. 2 shows the distributions of these entries across patient records. Fig. 3 quantifies the difference between the vital sign and laboratory value distributions between hospital systems using Jensen-Shannon divergence. Note that most clinical variables have similar distributions across hospital systems.

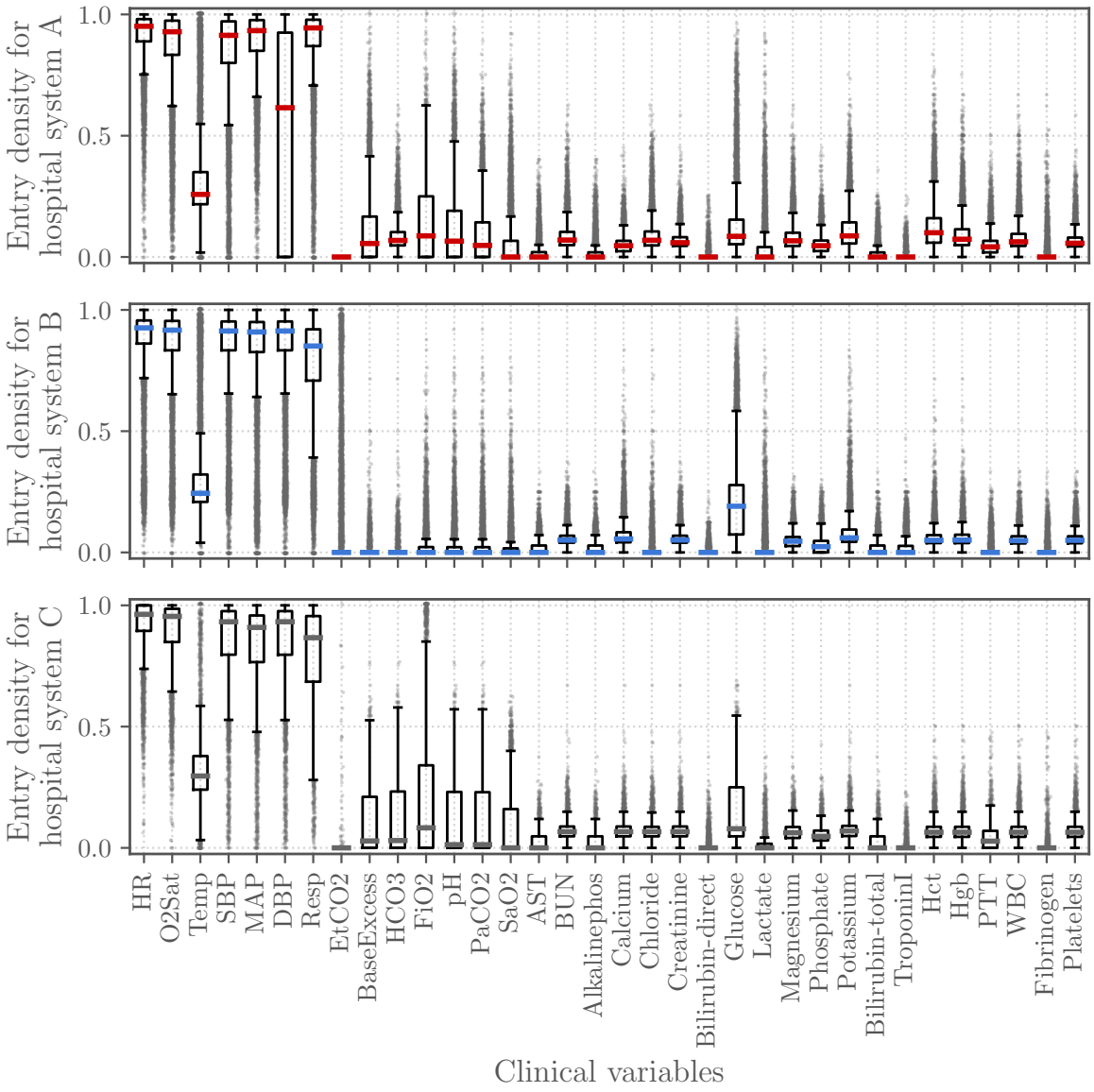


Figure 1: Densities of vital sign (rows 1-8) and laboratory value (rows 9-34) entries (fraction of non-empty entries) in the shared and hidden datasets for hospital systems A, B, and C.

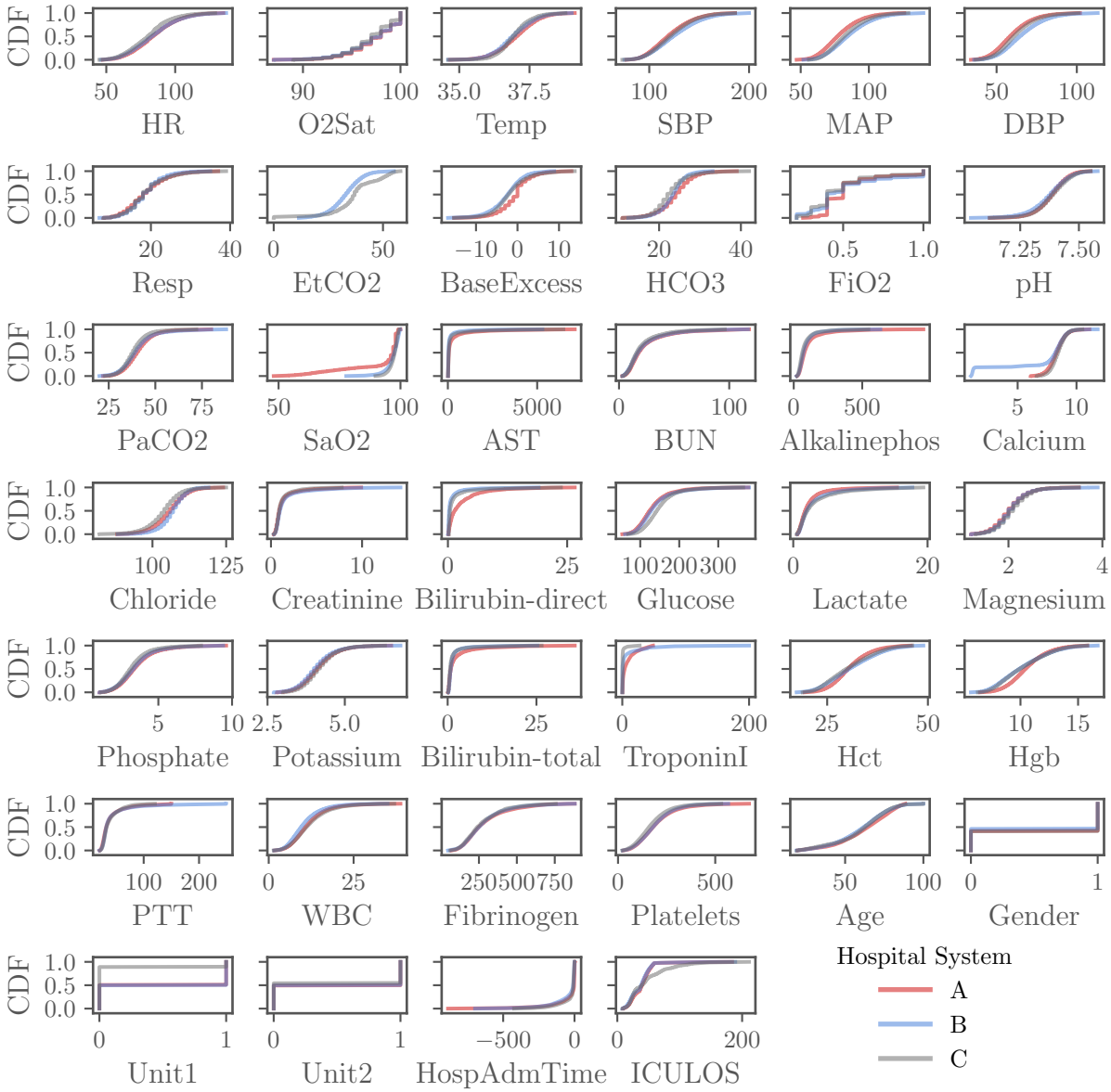


Figure 2: Distributions of vital sign (rows 1-8), laboratory value (rows 9-34), and demographic (rows 35-40) entries in the shared and hidden datasets for hospital systems A, B, and C; values below 0.5% or above 99.5% of each distribution not shown.

Hospital system	A	B
Number of patients	20,336	20,000
Number of septic patients	1,790	1,142
Sepsis prevalence	8.8%	5.7%
Number of rows	739,663	684,508
Number of entries	5,536,849	4,950,064
Density of entries	20.6%	19.1%

Table 2: Summary of vital sign (rows 1-8 in the data set) and laboratory value (rows 9-34) data in the shared datasets for hospital systems A and B.

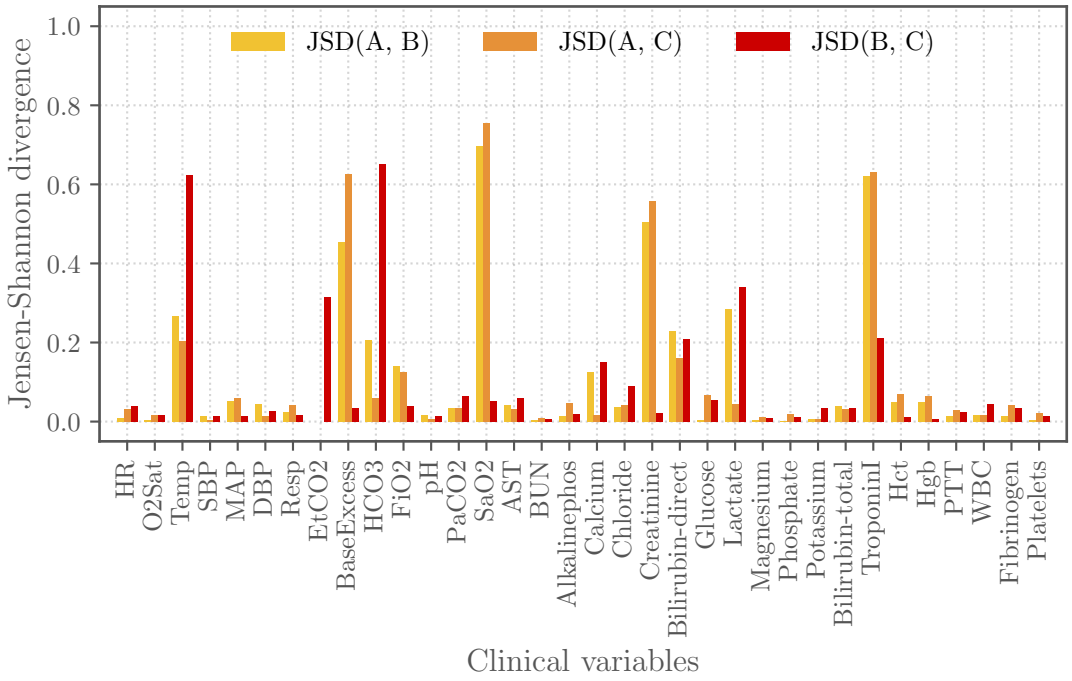


Figure 3: Jensen-Shannon divergence of vital sign (rows 1-8) and laboratory value (rows 9-34) distributions in the shared and hidden datasets for hospital systems A, B, C.

2.3 Challenge Scoring

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

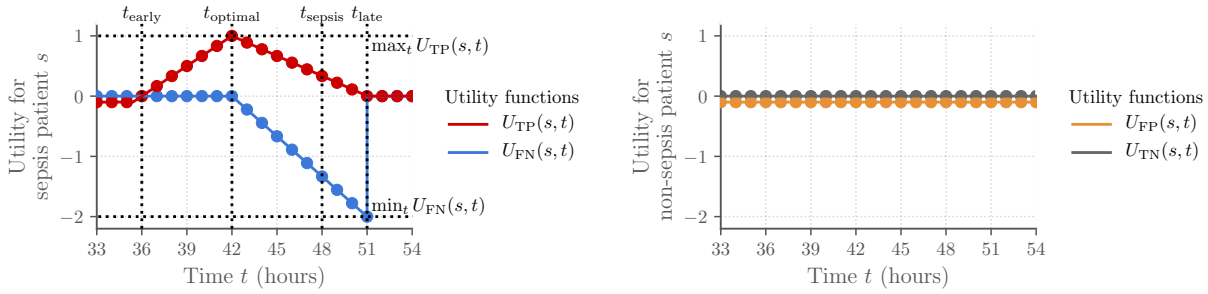
Each algorithm made a binary sepsis prediction for each hourly time window in each patient record. To evaluate each algorithm, we first defined a score for each prediction and then aggregated these scores over all hourly time windows and all patient records.

Given an algorithm's prediction for an hourly time window t in a patient record s, we define a score

$$U(s,t) = \begin{cases} U_{\mathrm{TP}}(s,t), & \text{positive prediction at time } t \text{ for a septic patient } s, \\ U_{\mathrm{FP}}(s,t), & \text{positive prediction at time } t \text{ for a non-septic patient } s, \\ U_{\mathrm{FN}}(s,t), & \text{negative prediction at time } t \text{ for a septic patient } s, \\ U_{\mathrm{TN}}(s,t), & \text{negative prediction at time } t \text{ for a non-septic patient } s, \end{cases} \tag{1}$$

where $U_{\mathrm{TP}}(s,t)$, $U_{\mathrm{FP}}(s,t)$, $U_{\mathrm{FN}}(s,t)$, and $U_{\mathrm{TN}}(s,t)$ are illustrated in Fig. 4a for an example septic patient and in Fig. 4b for an example non-septic patient. These scores were chosen to reflect the broad clinical realities of sepsis detection and treatment, and the actual utility values and time points in (1) and Fig. 4 can be chosen to capture the specific preferences or trade-offs of any particular hospital system.

For patients that become septic during their ICU stay, early sepsis detection tends to be beneficial. Therefore, sepsis predictions in septic patients that were at least 12 hours before and at most 3 hours after the onset time t_{sepsis} of sepsis were rewarded with a maximum reward at 6 hours before t_{sepsis}, and sepsis predictions that are more than 12 hours before



(a) Clinical utility for septic patients.

(b) Clinical utility for non-septic patients.

Figure 4: Diagrams of utility of positive and negative predictions for sepsis and non-septic patients; the time $t_{\text{sepsis}} = 48$ of sepsis onset is given as an example.

t_{sepsis} were slightly penalized. Similarly, for patients that become septic during their ICU stay, very early predictions may be implausible or unhelpful, and late or missed septic predictions are generally harmful. Therefore, sepsis predictions in septic patients that were more than 12 hours before t_{sepsis} were slightly penalized, and non-sepsis predictions that were less than 6 hours before t_{sepsis} were increasingly penalized.

For patients that do not become septic during their ICU stay, sepsis predictions contribute to alarm fatigue and lower confidence in algorithms, antibiotic overuse, and overall poor allocation of hospital attention and resources. Therefore, sepsis predictions in non-septic patients were slightly penalized. Similarly, non-sepsis predictions in non-septic patients were neither rewarded nor penalized.

Given an algorithm's predictions for all hourly time windows $T(s)$ in each patient record s, we define total score for an algorithm as the sum

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

over all predictions. For easier interpretability, we normalize (2) so that the optimal algorithm with the highest possible score receives a normalized score of 1 and a completely inactive algorithm that only makes non-sepsis predictions receives a normalized score of 0,

i.e.,

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

Each algorithm received a score from (3), and the algorithm with the highest value of (3) on the full sequestered dataset from hospital systems A, B, and C won the Challenge.

2.4 Challenge Submissions

Challenge participants submitted their algorithms for evaluation on the sequestered data. This strategy encouraged reproducibility and gave participants the ability to validate their algorithms on real-world datasets.

Each team was allowed a total of 5 scored entries during an unofficial phase of the Challenge from February 8, 2019 to April 14, 2019. This phase allowed for beta testing and socialization of the submission system, rules, and scoring mechanism, and teams were required to submit at least one entry during the unofficial phase for Challenge eligibility. Subsequently, each eligible team was allowed a total of 10 scored entries during the official phase of the Challenge from April 25, 2019 to August 25, 2019. This phase allowed teams to submit their models for evaluation on test data from hospital system A; scoring on the full hidden test data occurred only after the official phase at the end of the Challenge. This limit also improved the tractability of the Challenge. Since we did not heavily restrict the languages and libraries that teams could use, many teams required technical support for their submissions.

The submission system relied on containers that were orchestrated, as pipelines, on the Google Cloud Platform; Fig. 5 illustrates this system. A container is a standard unit of software that packages code and its dependencies so that the application runs readily and reliably in different computing environments. For the Challenge, we used the Docker containerization environment. Participants packaged their entries and uploaded them to a GitHub reposi-

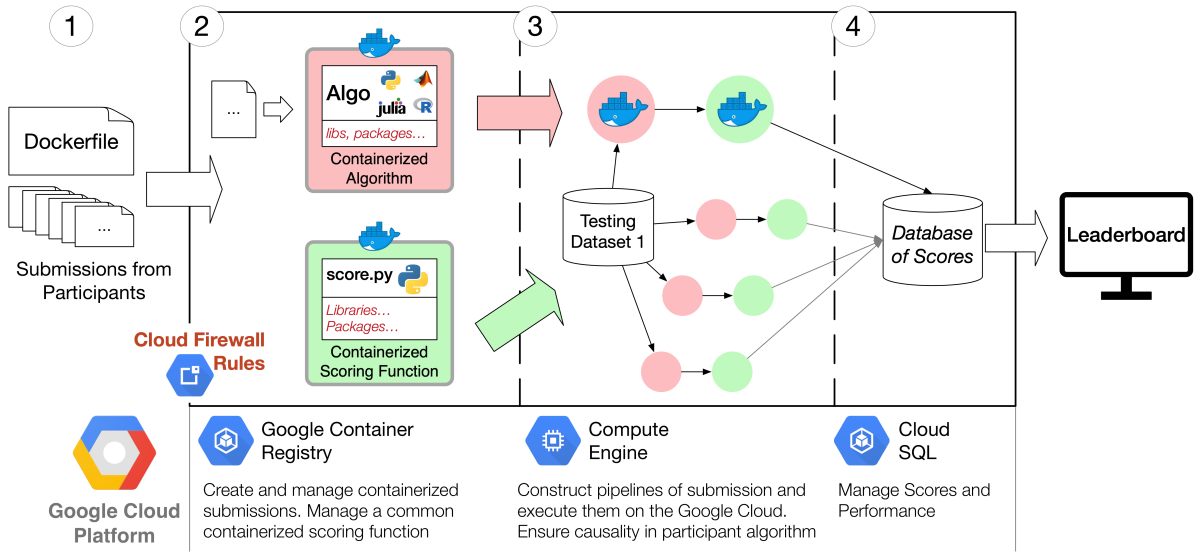


Figure 5: 1. Participants submitted Dockerfiles that contain recipes for creating containerized versions of their algorithm. 2. The submission system converted these Dockerfiles to containers and submitted them to the Google Container Registry. 3. An orchestration system, Cromwell, was used to retrieve the algorithm container and the containerized scoring function from the container registry and orchestrated them on the Google Compute Engine. The two containers ran as a pipeline. Cromwell and the submission system interfaced with the Compute Engine to launch the pipeline with the sequestered test dataset, and multiple pipelines could be executed in parallel. 4. The scores from the various pipelines were saved to a cloud database which drove a publicly accessible leaderboard.

tory, which was shared privately with the Challenge organizers. For each submission, the submission system cloned the repository, created a pipeline that consisted of the entry and our scoring function, and launched this pipeline on Google Cloud. This system allowed us to score multiple entries in parallel. During the unofficial and official phases of the Challenge, we processed over a thousand submissions in Julia, MATLAB, Python, and R from over a hundred participants.

Each entry was run in a virtual machine (VM) with 2 CPUs and 12 GB of RAM, and each entry was allowed 24 hours of run time on each hidden test set. The submission system orchestrator, Cromwell, typically requested a `n2-highmem-2` machine type on Google Cloud.

2.5 Implementations of Evaluation Metric and Baseline Model

To provide a baseline model, we trained a Weibull-Cox regression model and provided open-source implementations of this model in Julia, MATLAB, Python, and R. These implementations also served as examples of how to devise a working prediction algorithm in each language that we accepted for the Challenge. We also provided open-source implementations of our clinical utility-based scoring function. The code is available online at `https://github.com/physionetchallenges`.

2.6 Analysis of Entry Independence, Collusion and Plagiarism

After the conclusion of each Challenge, we frequently build a meta-algorithm from the final entries that are weighted by their independence; agreement between highly similar algorithms can suggest a false consensus of predictions. To increase the independence of algorithms, we therefore prohibited teams from collaboration at any point of the Challenge. Specifically, we note:

- Multiple teams from a single entity (such as a company, university, or department) were permitted as long as the teams were truly independent and did not share team members, code, or ideas at any point. Multiple teams from the same research group or unit within a company were not allowed because we did not believe that true independence between teams could be maintained when team members may frequently interact.

- New team members could join as long as they had not previously been involved with another team or had communicated with a team member from another team concerning this year's Challenge.

- Teams could use public code if it had been posted before the competition. Members

of teams were not allowed to publicly post code during the competition or use another competitor's code that was posted during the competition whether or not it was intentionally made public.

- Members of teams were not allowed to publicly post information describing their methods or give a talk outside of their own research group at any point during the competition that revealed the methods they have employed or planned to employ in the Challenge. Members of teams were allowed to present or publish on methods on other data as long as they did not indicate that they planned to apply it to Challenge data until after the competition.

- Members of teams were required to use the same team name and email address throughout the course of the competition, including for abstract submissions to the public forum at which they defended their work, i.e., at Computing in Cardiology.

Although the rules of the Challenge strictly prohibited teams from more than 10 scored entries during the official phase of the Challenge, several entries from apparently different teams achieved exactly the same score. An investigation of their submissions showed strong similarities between these teams, which, when questioned, either did not reply or claimed not to have colluded. By examining associations between email addresses, team names, and GitHub repositories, we were able to identify several prohibited collaborations. Fig. 6 illustrates associations between email addresses, team names, and GitHub users from Challenge submissions, where each team was expected to have only one email address, team name, and GitHub user. Some associations with multiple email addresses, team names, and/or GitHub users indicated prohibited collaborations and resulted in disqualifications.

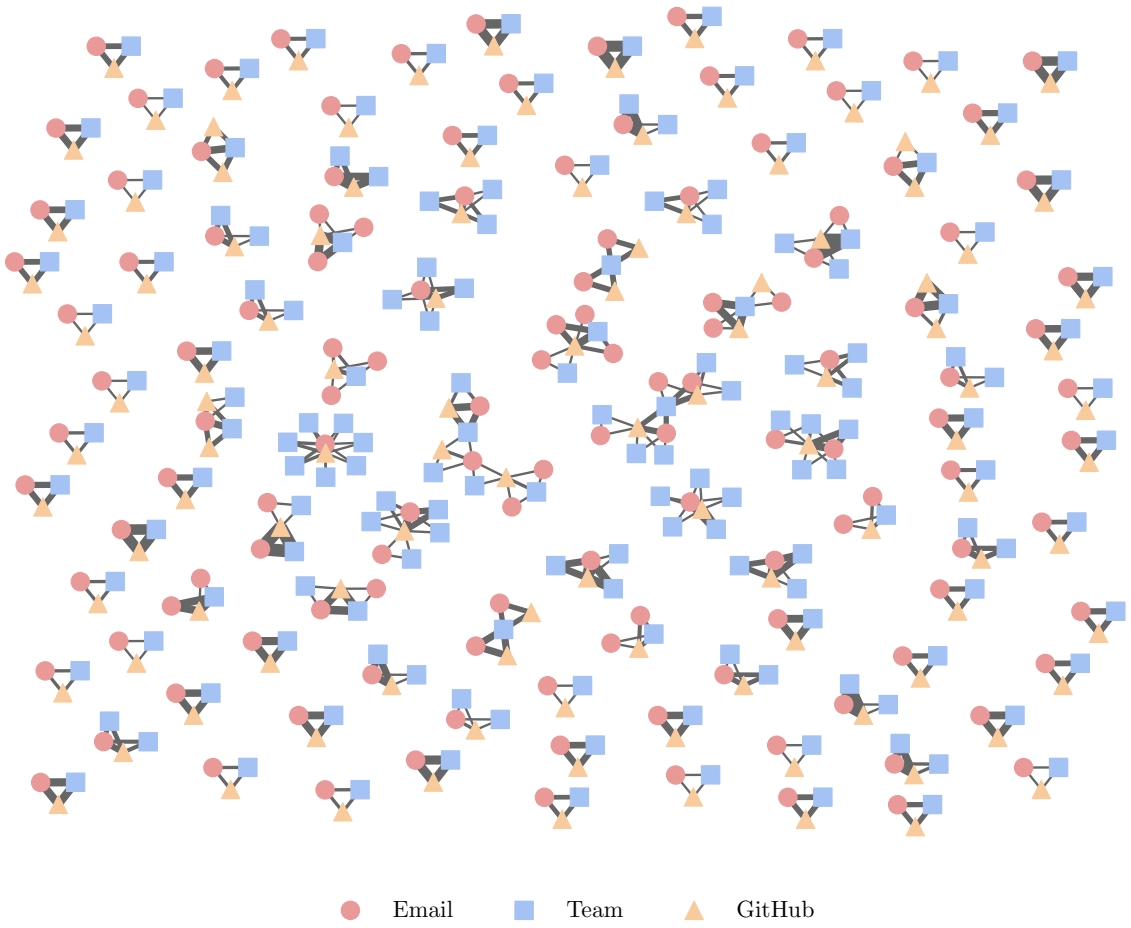


Figure 6: Collaboration between teams: vertices indicate email addresses, team names, and GitHub users, and edges indicate that a participant included an email address, team name and GitHub user as part of the same submission; multi-edges indicate multiple submissions and are shown with thicker line weights.

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 3: 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.

2.7 Results

A total of 104 teams from academia and industry submitted a total of 853 entries during the official phase of the Challenge; of these, 88 distinct teams with a total of 430 entries were able to be scored. Recall that each team received training data and labels for hospital systems A and B but not for hospital system C. Each successful entry received scores on the test data for hospital system A during the unofficial and official phases of the Challenge, and each team nominated its favorite successful entry for evaluation on the full test data containing patient records from hospital systems A, B, and C. Table 3 summarizes the teams with the highest-scoring entries.

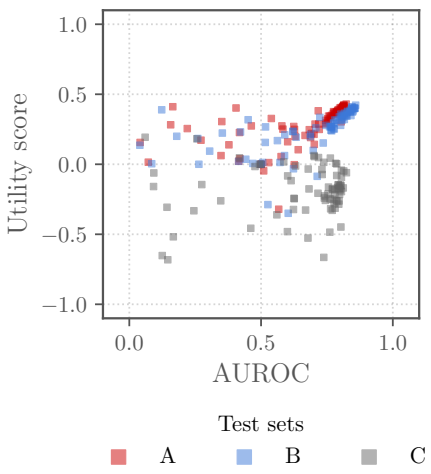
By curating clinical data from multiple hospital systems and sharing different amounts of data and information from these systems, we demonstrated that algorithms generally performed much better in two hospital systems for which we provided training data than a third hospital system for which we provided no training data.

While algorithms that performed well by one evaluation metric might be expected to perform well by another metric, we saw that this was generally not the case for traditional

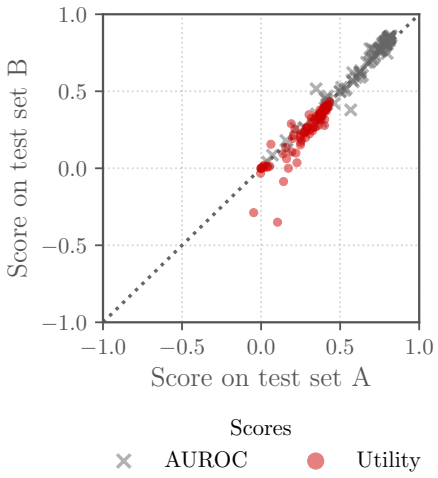
evaluation metrics and the clinical utility score that we devised for the Challenge. Fig. 7a compares each algorithm's AUROC with its utility score on the test sets from each of the hospital systems. AUROC and utility scores are positively correlated on test sets A and B (Spearman rank correlation coefficients $\rho = 0.791$ and $\rho = 0.839$, respectively). These scores are poorly correlated on test set C (Spearman rank correlation coefficient $\rho = 0.054$), which corresponds to the hospital system for which participants did not receive training data. Moreover, even on test sets A and B, algorithms with high utility scores did not necessarily have high AUROC scores, demonstrating that traditional evaluation metrics do not necessarily capture the clinical utility of predictions.

Moreover, the choice of evaluation metric influenced how transferable algorithms appeared to be across hospital systems. Figs. 7b, 7c, 7d compare each algorithm's AUROC or utility score on test sets from different hospital systems. While AUROC scores are strongly correlated for each pair of hospital systems (Spearman rank correlation coefficients $\rho = 0.973$ for hospital systems A and B, $\rho = 0.938$ for hospital systems A and C, and $\rho = 0.947$ for hospital systems B and C), this is not true for utility scores. Utility scores are strongly correlated between the two hospital systems for which we provided training data (Spearman rank correlation coefficient $\rho = 0.949$ for test sets A and B), but they are poorly correlated with the third hospital system for which we did not provide training data (Spearman rank correlation coefficient $\rho = -0.033$ and $\rho = 0.013$ for hospital systems A and B, respectively, with hospital system C). Fig. 7e further shows that the methods with the highest scores on data from hospital systems with shared training databases were not necessarily the methods with the highest scores on the hidden database from a separate hospital system.

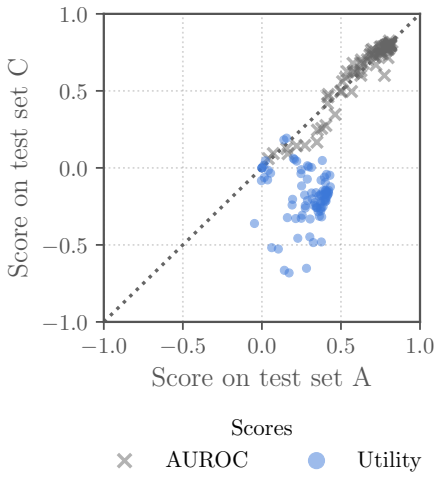
Our use of clinical data from multiple hospital systems and our application of a clinical utility-based evaluation metric provided a more nuanced view of predictive generalizability than results on one system with traditional evaluation metrics would present.



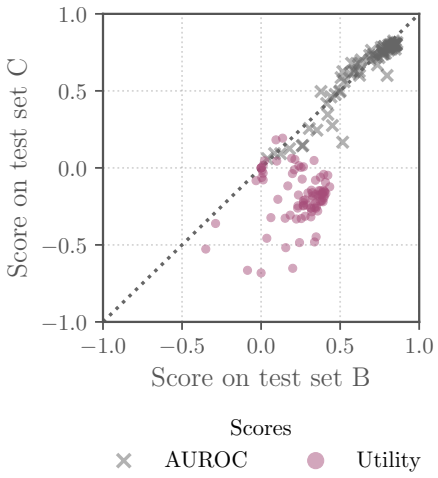
(a) Comparison of each algorithm's AUROC and utility scores on test sets A, B, and C.



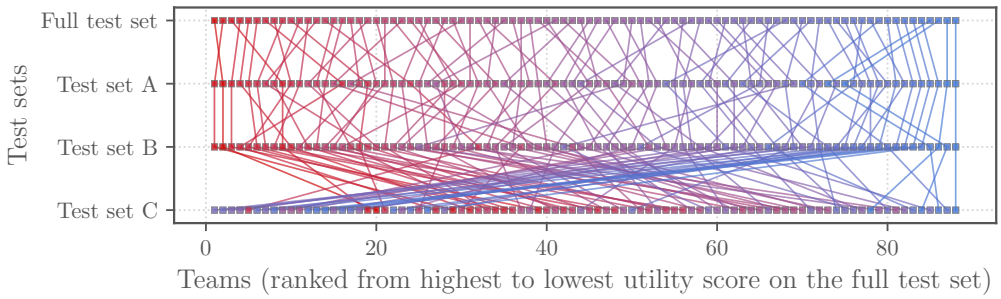
(b) Comparison of each algorithm's AUROC and utility scores on test sets A and B.



(c) Comparison of each algorithm's AUROC and utility scores on test sets A and C.



(d) Comparison of each algorithm's AUROC and utility scores on test sets B and C.



(e) Ranked performance of the final algorithms on test sets A, B, and C. Red indicates a high overall ranking across all three databases, and blue indicates a low overall ranking. Lines from top to bottom indicate how the individual algorithm ranking changed when considering the performance on each database. Algorithms that performed well on test sets A and B generally performed relatively poorly on test set C.

Figure 7: Comparison of each algorithm's AUROC and utility scores on test data from hospital systems A, B, and C, where we shared training data for hospital systems A and B but not for hospital system C.

3 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 over 60,000 patient records from three hospital systems, with two shared publicly and one remaining hidden. By posting two databases publicly, 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 also proposed and used a novel evaluation metric that captures the clinical utility of early sepsis detection, weighted by the relative "earliness" or "lateness" of each prediction. We suggest that this metric should be considered for wider adoption in clinical care because it does not suffer from many of the problems of F-measures (and related metrics such as accuracy, sensitivity, and positive predictive value) or standard area under the curve metrics (such as AUROC and AUPRC), which either assume a one-shot decision or no decision threshold, respectively. In particular, this novel evaluation metric shows that algorithms that perform well in one hospital system may perform poorly in another.

A third novelty in this Challenge is the development of graphical and analytical approaches to measure the similarity between entries between supposedly independent Challenge teams. We identified and disqualified teams that appeared to be highly related to each other and did not provide satisfactory explanations of these relationships.

We received 853 entries from 104 participants in academia and industry, providing a diverse view of algorithmic approaches to early sepsis detection. Combined, these efforts provide a more complete picture of how algorithms can provide early sepsis predictions. A subsequent analysis of the best performing and most interesting algorithms submitted to the Challenge will combine the strengths of different approaches to push the boundaries of

automated approaches to early sepsis prediction.

References

[1] Christopher W. Seymour, Vincent X. Liu, Theodore J. Iwashyna, Frank M. Brunkhorst, Thomas D. Rea, André Scherag, Gordon Rubenfeld, Jeremy M. Kahn, Manu Shankar-Hari, Mervyn Singer, 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*, 315(8):762–774, 2016.

[2] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *Journal of the American Medical Association*, 315(8):801–810, 2016.

[3] Manu Shankar-Hari, Gary S Phillips, Mitchell L Levy, Christopher W Seymour, Vincent X Liu, Clifford S. Deutschman, Derek C. Angus, Gordon D. Rubenfeld, and Mervyn Singer. 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*, 315(8):775–787, 2016.

[4] Centers for Disease Control and Prevention. Sepsis. `https://www.cdc.gov/sepsis/datareports/index.html`, 23 August 2016. [Online; accessed 1 February 2019].

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

[6] Carly J. Paoli, Mark A. Reynolds, Meenal Sinha, Matthew Gitlin, and Elliott Crouser. Epidemiology and Costs of Sepsis in the United States—An Analysis Based on Timing of Diagnosis and Severity Level. *Critical Care Medicine*, 46(12):1889, 2018.

[7] Derek C. Angus, Walter T. Linde-Zwirble, Jeffrey Lidicker, Gilles Clermont, Joseph Carcillo, and Michael R. Pinsky. Epidemiology of severe sepsis in the United States: analysis of incidence, outcome, and associated costs of care. *Critical Care Medicine*, 29(7):1303–1310, 2001.

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

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

[10] Jeremy Stoller, Laura Halpin, Matthew Weis, Brett Aplin, Weikai Qu, Claudiu Georgescu, and Munier Nazzal. Epidemiology of severe sepsis: 2008-2012. *Journal of Critical Care*, 31(1):58–62, 2016.

[11] Christopher W. Seymour, Foster Gesten, Hallie C. Prescott, Marcus E. Friedrich, Theodore J. Iwashyna, Gary S. Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M. Terry, and Mitchell M. Levy. Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine*, 376(23):2235–2244, 2017.

[12] Vincent X. Liu, Vikram Fielding-Singh, John D. Greene, Jennifer M. Baker, Theodore J. Iwashyna, Jay Bhattacharya, and Gabriel J. Escobar. The Timing of Early Antibiotics

and Hospital Mortality in Sepsis. *Am J Respir Crit Care Med*, 196(7):856–863, October 2017.

[13] Anand Kumar, Daniel Roberts, Kenneth E. Wood, Bruce Light, Joseph E. Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Sergio Zanotti, Leo Taiberg, et al. Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine*, 34(6):1589–1596, 2006.

[14] Hallie C. Prescott and Theodore J. Iwashyna. Improving Sepsis Treatment by Embracing Diagnostic Uncertainty. *Annals of the American Thoracic Society*, 16(4):426–429, 2019.

[15] Ary L. Goldberger, Luis A.N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation*, 101(23):e215–e220, 2000.

[16] Katharine E. Henry, David N. Hager, Peter J. Pronovost, and Suchi Saria. A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*, 7(299):299ra122–299ra122, 2015.

[17] Shamim Nemati, Andre Holder, Fereshteh Razmi, Matthew D Stanley, Gari D Clifford, and Timothy G Buchman. An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical Care Medicine*, 46(4):547–553, 2018.

[18] Li-Fang Cheng, Niranjani Prasad, and Barbara E. Engelhardt. An optimal policy for patient laboratory tests in intensive care units. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 24, pages 320–331. World Scientific, 2019.

[19] Anton Aboukhalil, Larry Nielsen, Mohammed Saeed, Roger G. Mark, and Gari D. Clifford. Reducing false alarm rates for critical arrhythmias using the arterial blood pressure waveform. *Journal of Biomedical Informatics*, 41(3):442–451, 2008.