Time-specific Metalearners for the Early Prediction of Sepsis

Marcus Vollmer[1], Christian F. Luz[2], Philipp Sodmann[1], Bhanu Sinha[2], Sven-Olaf Kuhn[3]

[1] Institute of Bioinformatics, University Medicine Greifswald, Germany
[2] Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, The Netherlands
[3] Department of Anesthesiology and Intensive Care Medicine, University Medicine Greifswald, Germany

Abstract

Motivation: *This contribution relates to the* PhysioNet/CinC Challenge 2019 *on real-time early detection of sepsis from bedside monitoring and laboratory parameters. Accounting for complex clinical dynamics in sepsis patients while aiming at an automated analysis of hourly (non-)validated data is challenging. The algorithm has to deal with imprecise, incorrect and incomplete data in addition to being time aware.*

Methods: *We cleaned the data and build features (e.g. assumed presence of ventilation) based on rolling windows and included several clinical scores, such as Shock Index, qSOFA, SOFA, SIRS, NEWS, cNEWS. We aimed to build time-specific stacked ensembles and a non-specific XGBoost baselearner to predict sepsis 6 hours prior to the sepsis onset. The models were trained on a triple split of 40,336 ICU patients taken from the freely available training sets of the* 2019 PhysioNet/CinC Challenge*. 60% was used for training, 20% for hyperparameter optimization and validation, and 20% for independent testing. The performance was evaluated using task-specific utility functions that rewards predictions between 12 hours before and 3 hours after the sepsis onset (as defined by the Sepsis-3 guidelines) with a defined optimum at 6 hours before. Furthermore, variable importance was assessed.*

Results: *A normalized utility score of* 0.394 *can be reported for an non-specific XGBoost model for the independent test set. The threshold selection was more vague in time-specific meta-learners that leads to an inferior performance. Most important variables include the assumed presence of ventilation, white blood cells, partial thromboplastin time, blood urea nitrogen and rolling quantiles of the temperature. Partial SOFA-scores, cNEWS, and the Shock Index showed importance only in the ICU admission phase.*

1. Introduction

Time is life – this mantra of emergency medicine also applies to one of the most dangerous clinical situations in critical care: sepsis. The Third International Consensus Definitions for sepsis and septic shock (Sepsis-3) defined sepsis as life-threatening organ dysfunction caused by a dysregulated host response to infection. Septic shock is a subset of sepsis characterized by persistent arterial hypotension requiring vasopressor support despite adequate fluid resuscitation. Furthermore, perfusion abnormalities, such as oliguria, reduced peripheral perfusion, and altered mental status occur [1]. The appearance of sepsis is highly individual and demonstrated by the origin of the infection, and different predisposing factors like underlying genetic variation and immune response state. Despite overall medical progress and standardized guidelines promoting immediate actions when sepsis is suspected, diagnosis of sepsis in critically ill patients is challenging and mortality remains high [2]. Dutch intensive care units (ICU) reports bloodstream infections (often causing sepsis) to be the fourth most common reason for ICU admission with a mortality of 18.5, 32.3, and 40.0% during admission and three and twelve months after admission respectively [3]. ICUs are among the most data-intense environments in hospitals. Routinely available data such as vital signs and laboratory parameters have been studied for the (early) detection of septic patients since decades. Systemic inflammatory response syndrome (SIRS) criteria or sequential organ failure assessment (SOFA) scores are examples of such derived approaches used in patient monitoring and clinical decision making. However, the applicability is limited due to the trade-off with simplicity. Nowadays, modern machine learning algorithms have the potential to leverage routinely available data to the maximum and support clinicians in (early) detection of sepsis in critically ill patients. Recent studies have explored several machine learning approaches in tackling this challenge [4–7]. Despite different sepsis definitions and target outcomes such as clinical definition based on the sepsis-3 guideline, ICD codes for sepsis, bacteremia, or mortality, decently performing machine learning models have been described. Scores derived from Random forest, linear regression, and especially long short-term memory models have demonstrated to largely outperform traditional scores

(e.g. SIRS, SOFA) while using traditionally available data like vital signs and laboratory parameters [4–7].

Our approach has a special medical focus on data preprocessing, data cleaning, and outlier detection. New variables were generated based on clinical experience and available data (e.g. presence of ventilation or oxygen partial pressure estimates). Clinical scores per time point and rolling window were defined and incorporated in the preprocessing steps. Imputation methods were used that most closely mimic clinical reasoning.

2. Data Screening and Cleaning

Data is based on the training set of the PhysioNet/Computing in Cardiology Challenge 2019. Our aim was to optimize a specific utility function with a reward for predicting sepsis in a time window of 12 hours before and 3 hours after given sepsis onset and specific penalties for false negative and false positive predictions, see full challenge description in [8]. The challenge dataset consists of 1,552,210 data points from 40,336 patients admitted to a medical and a surgical ICU in the USA. Included are the hour post ICU admission, the marker for sepsis and hourly measured vitals, lab values, and basic demographics.

First, we had a look into the length of ICU stay and the number of patients with available data at a specific time after ICU admission. In Figure 1 we noticed the hospital specific discharge policy leading to an immense drop of patient data after $36\,\mathrm{h}$ and $60\,\mathrm{h}$ post ICU admission. Because of the prolonged stay of critically ill patients, the proportion of sepsis rises to approximately $12.5\,\%$ after $60\,\mathrm{h}$. At this time, the data availability changes and sepsis definition turns from home- or hospital-acquired to ICU-acquired sepsis. Second, vitals and lab values were screened for physiological plausibility and 2263 values (mainly within blood pressure variables, respiration rate and oxygen levels) were removed from the training data set. Next, we tabulated data availability of the 12,036,860 remaining values and identified the rhythm of measurements for all variables, see Table 1. The table is showing the number of data points without gaps (hourly measured: n0), with exactly one missing value between the last observation (n1) and relative cumulative sums (q0, q1, q2). Temperature

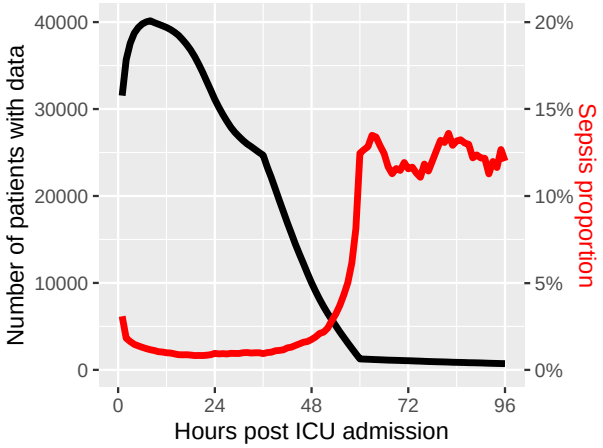

Figure 1. Data availability and proportion of sepsis.

for instance was measured in only $34\,\%$ of all hourly data points, whereas heart rate was the most frequently measured variable ($90\,\%$) in this clinical setting.

3. Feature Engineering

We implemented rolling windows of 6, 12, 24 and 48 hours for frequently repeated features, such as heart rate, oxygen saturation, temperature, systolic/diastolic/mean atrial blood pressure, respiration rate and serum glucose, to compute quantiles, quantile ranges, and differences and quotients to the actual value. Quantiles $(0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95)$ were chosen to represent the course of a disease by excluding outliers. We used the last observation carried forward method to copy the last available lab and vital value to the actual date if new data is missing. This approach represents the medical perspective of decision making and lab values from blood samples are usually measured with varying frequency. Next, we made missingness explicit by introducing binary variables to indicate whether the values were carry-forwarded. Additionally, we introduced numerical variables showing the up-to-dateness of the given value (0 means newly measured, 6 means measured 6 hours ago) so that machine learning models are able to learn the relevance of out-dated variables. Table 2 shows some features described above for the mean atrial pressure (map) measured in the first 10 hours of a patient. Empty cells indi-

Table 1. Number of gaps/missing data between observation for a subset of variables.

Variables	n0	n1	n2	n3	q0	q1	q2
Heart rate (hr)	1398740	87915	8904	6524	0.90	0.96	0.96
Oxygen saturation (o2sat)	1349202	94193	13086	9352	0.87	0.93	0.94
Body temperature (temp)	525111	56587	49411	160226	0.34	0.37	0.41
Systolic blood pressure (sbp)	1325945	97290	11865	8200	0.85	0.92	0.92
Mean atrial pressure (map)	1358496	99527	11949	6842	0.88	0.94	0.95
Diastolic blood pressure (dbp)	1065282	68661	9376	7266	0.69	0.73	0.74
Respiration rate (resp)	1313516	100718	15486	10159	0.85	0.91	0.92
End tidal carbon dioxide (et_co2)	57636	3407	562	367	0.04	0.04	0.04

Table 2. Feature engineering on mean atrial pressure (map) of a patient subset.

ICU length of stay (ICULOS)	1	2	3	4	5	6	7	8	9	10
Mean atrial pressure (map_raw)		75.3	86.0		91.3		77.0	76.3	88.3	87.3
Carry-forwarded values (map_LOCF)		75.3	86.0	86.0	91.3	91.3	77.0	76.3	88.3	87.3
Missingness (map_miss)	T	F	F	T	F	T	F	F	F	F
Missing value (map_miss_val)	1	0	0	1	0	1	0	0	0	0
50% quantile of the last 6h (map_roll.t6.p50)		75.3	80.7	80.7	86.0	86.0	81.5	81.5	82.7	87.3
75% quantile of the last 6h(map_roll.t6.p75)		75.3	83.3	83.3	88.7	88.7	87.3	87.3	89.1	88.3

cating missingness of data which is explicitly tracked by the 'miss' variables. The robust variable generation is followed by the computation of assumend presence of ventilation (equaling to available etCO2 measurements) and the estimation of partial pressure of oxygen (PaO2) from oxygen saturation (saO2). Furthermore, various clinical scores were calculated:

- **ShockIndex** (hr/sbp)
- **qSOFA** (sbp and resp)
- **SOFA** and partial SOFA scores (respiration, renal function, platelets, liver function, sofa_renal, sofa_plate, mean arterial pressure), SOFA from worst 24h partial scores
- **SIRS** criteria [9], worst 24h SIRS score, SIRS criteria with hard temperature thresholds
- **NEWS (National Early Warning Score)** and partial NEWS scores (respiration, oxygen saturation, systolic blood pressure, pulse rate, temperature)
- **cNEWS** [10] uses linear regression (gender, age, NEWS, log(resp), temp, log(sbp), log(dpb), log(hr), o2sat, o2support)
- **Rolling versions using robust measures:**
  - qSOFA_t6 uses 25% and 75% quantiles of last 6h
  - shockIndex_t6 uses 25% and 75% quantiles of last 6h
  - SIRS_t24 and partial scores uses 25%, 75% quantiles for temperature and 90% quantiles of the last 24h for heart rate and respiratory rate
  - NEWS_t6 uses 50% quantiles of respiratory rate, heart rate and systolic blood pressure of the last 6h

The final size of the generated dataset was $1,552,210 \times 427$ and the patient-wise computation of rolling variables on previous data made it possible to use just a single row for sepsis prediction.

4. Automated Machine Learning

Considering the changes in data availability and in sepsis prevalence during the course of the ICU stay, we aimed to adapt to the changing demands and trained a time-specific ensemble learner (metalearner). We also compared the predictions with an XGBoost base-learner trained on the entire dataset (with opportunity to use the ICULOS as a predictor). We split the data and used $60\,\%$ of all patients for training, $20\,\%$ for validation, and $20\,\%$ for independent testing of the derived models. The validation set was used for hyperparameter optimization and the computation of a threshold to transform the sepsis score into binary classes (sepsis/non-sepsis). Model building was performed in R using the H2O package [11] to train machine (XGBoost, GBM, DRF, GLM) and deep learning models and to solve the binary classification task (non-sepsis/sepsis). Sepsis was defined as a union of pre-sepsis data points (range of data points 12 hours prior to 6 hours prior to the sepsis onset), sepsis data points (from 6 hours prior to 3 hours post sepsis onset), and post-sepsis data points (3 hours post sepsis onset and later). We used 5-fold cross-validation, user-specific class sampling factors (0.1 for non-sepsis, 2.0 for sepsis) and logloss as the stopping metric within the fitting processes. Furthermore, a stacked ensemble (SE) was build to further improve the predictability. Figure 2 illustrates our workflow.

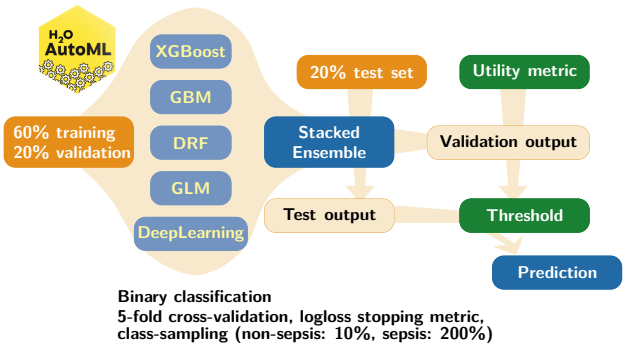

Figure 2. Scheme illustrating the workflow.

5. Evaluation of Predictions & Results

We evaluated the threshold method by computing the normalized utility values U_{norm} (see [8]) at each possible threshold in the training, validation and test set. The threshold was selected at the sepsis score with the maximal U_{norm} in the validation set and the final scoring was extracted from the test set as illustrated in Figure 3. Using the XGBoost base-learner based on the complete training set, we were able to reach a normalized utility value of 0.394 at a threshold of 0.03496. The boxplot in Figure 3 shows the sepsis scores for all dates in the test set on logarithmic scale. The suggested threshold would identify more than $60\,\%$ of all sepsis data points and no statistical difference can be found in the scores of PreSepsis, Sepsis and PostSepsis. Moreover, almost $90\,\%$ of non-sepsis data were correctly classified.

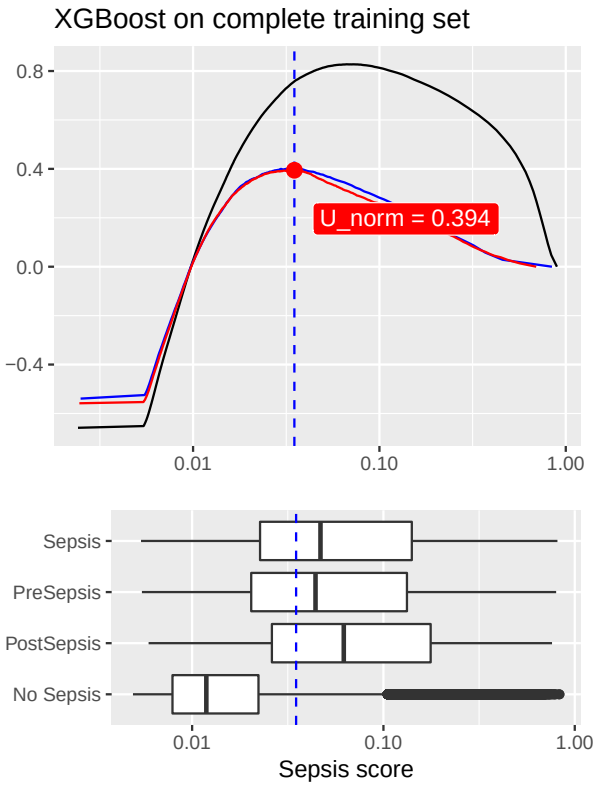

Figure 3. Thresholding maximizes the normalized utility function and score distribution of patients prediction in the test test.

By selecting specific ICULOS dates, we evaluated the time-specific normalized utility scores to compare the time-specific SE meta-learners with the non-specific XGBoost learner. We observed that with the time-specific meta-learner the threshold selection is more vague and led to an inferior performance. We extracted the variable importance of related models and identified assumed presence of ventilation, white blood cells, partial thromboplastin time, blood urea nitrogen and rolling quantiles of the temperature to be amongst the TOP 15 predictors independently of the ICULOS. Important variables for predicting sepsis in the admission phase were partial SOFA-scores, cNEWS, and the shock index. cNEWS was top-ranked also till 18 h post admission.

References

[1] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche JD, Coopersmith CM, Hotchkiss RS, Levy MM, Marshall JC, Martin GS, Opal SM, Rubenfeld GD, van der Poll T, Vincent JL, Angus DC. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA 2016;315(8):801–10.

[2] Rhodes A, Evans LE, Alhazzani W, Levy MM, Antonelli M, Ferrer R, Kumar A, Sevransky JE, Sprung CL, Nunnally ME, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. Intensive care medicine 2017;43(3):304–377.

[3] Stichting NICE. Nationale intensive care evaluatie jaarboek 2016. Technical report, Stichting NICE, 2017.

[4] Taylor RA, Pare JR, Venkatesh AK, Mowafi H, Melnick ER, Fleischman W, Hall MK. Prediction of in-hospital mortality in emergency department patients with sepsis: A local big Data-Driven, machine learning approach. Acad Emerg Med 2016;23(3):269–278.

[5] Calvert JS, Price DA, Chettipally UK, Barton CW, Feldman MD, Hoffman JL, Jay M, Das R. A computational approach to early sepsis detection. Comput Biol Med 2016;74:69–73.

[6] Kam HJ, Kim HY. Learning representations for the early detection of sepsis with deep neural networks. Comput Biol Med 2017;89:248–255.

[7] Van Steenkiste T, Ruyssinck J, De Baets L, Decruyenaere J, De Turck F, Ongenae F, Dhaene T. Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks. Artificial Intelligence in Medicine 2019;97:38–43.

[8] Matthew Reyna Supreeth Prajwal Shashikumar BMPGASsnGC. Early Prediction of Sepsis from Clinical Data: the PhysioNet/Computing in Cardiology Challenge 2019. In Computing in Cardiology 2019. ISSN 2325-886X, 2019; 1–4.

[9] Rangel-Frausto MS, Pittet D, Costigan M, Hwang T, Davis CS, Wenzel RP. The natural history of the systemic inflammatory response syndrome (SIRS): a prospective study. Jama 1995;273(2):117–123.

[10] Faisal M, Richardson D, Scally AJ, Howes R, Beatson K, Speed K, Mohammed MA. Computer-aided national early warning score to predict the risk of sepsis following emergency medical admission to hospital: a model development and external validation study. CMAJ 2019;191(14):E382–E389.

[11] Aiello S, Eckstrand E, Fu A, Landry M, Aboyoun P. Machine learning with r and h2o. H2O booklet 2016;.

Address for correspondence:

Marcus Vollmer / marcus.vollmer@uni-greifswald.de
Institute of Bioinformatics / University Medicine Greifswald
Felix-Hausdorff-Str. 8 / 17475 Greifswald / Germany