Time-Specific Metalearners for the Early Prediction of Sepsis

Marcus Vollmer¹, Christian F Luz², Philipp Sodmann¹, Bhanu Sinha², Sven-Olaf Kuhn³

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: 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 aimed to build time-specific stacked ensembles and a non-specific XGBoost learner to predict sepsis 6 hours prior to the sepsis onset. The models were trained on a triple split of 40,336 ICU stays taken from the training sets of the 2019 PhysioNet/CinC Challenge. Data was cleaned and features were built based on rolling windows including several clinical scores and criteria, such as shock index, qSOFA, SOFA, SIRS, NEWS, cNEWS. Model performance was evaluated using task-specific utility functions. Furthermore, variable importance was assessed.

Results and conclusion: Although no official score was obtained in the Challenge as team Sepsis2G, we found normalized utility score of 0.394 for our non-specific XGBoost model on a held out subset of the training data. The threshold selection was displaced in time-specific metalearners leading to an inferior performance. Most important variables included the assumed presence of ventilation, white blood cell count, partial thromboplastin time, blood urea nitrogen and rolling quantiles of the temperature. Partial SOFA-scores, cNEWS, and the shock index showed major importance 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. Dutch intensive care units (ICU) report bloodstream infections (often causing sepsis) to be the fourth most common reason for ICU admission with a three month mortality rate of 32.3% [1]. The Third International Consensus Definitions for sepsis and septic shock (Sepsis-3) defined sepsis as “life-threatening organ dys-function 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 [2]. The clinical presentation of sepsis is highly diverrs as a consequence of different origins of the infection, and different predisposing factors such as 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 [3]. ICUs are among the most data-intense environments in hospitals. Routinely available data such as vital parameters and laboratory results 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 between simplicity and the heterogeneous nature of sepsis. Nowadays, modern machine learning algorithms have the potential to leverage routinely available data to the maximum and support clinicians in detection of sepsis in critically ill patients. Scores derived from Random forest, linear regression, and especially long short-term memory models have demonstrated to largely outperform traditional clinical scores (e.g. SIRS, SOFA) while using traditionally available data like vital parameters and laboratory results [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 for missing data were used that most closely mimic clinical reasoning.
2. Data Screening and Cleaning

Used data is based on the training set of the 2019 PhysioNet/Computing in Cardiology Challenge. 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 dataset consists of 1,552,210 data points from 40,336 patients admitted to medical and surgical ICUs at two US-American hospitals. This data comprised basic patient demographics, hourly measured vital parameters, laboratory results, the hour ICU admission, and a binary label of the presence of sepsis.

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 were assessed. In Figure 1 a hospital specific discharge policy leading to an immense drop of patient data after 36 h and 60 h post ICU admission could be identified. Due to the prolonged stay of critically ill patients, the proportion of sepsis increases to approximately 12.5 % after 60 h. At this time, the data availability changes and sepsis definition turns from community-acquired or hospital-acquired to ICU-acquired sepsis. Second, vital parameters and laboratory results were screened for physiological plausibility and 2263 values (mainly within blood pressure variables, respiration rate and oxygen levels) were removed from the data set. Next, data availability of the 12,036,860 remaining values were tabulated and the rhythm of measurements for all variables was assessed. Table 1 depicts the number of data points without gaps (hourly measured: n=0), with exactly one (n=1), two (n=2), or three (n=3) missing values between the last observation and relative cumulative sums in percent. Temperature for instance was measured hourly in only 34 % of all data points, whereas heart rate was the most frequently measured variable (90 %).

3. Feature Engineering

Rolling windows of 6, 12, 24 and 48 hours were implemented to compute quantiles, quantile ranges, and differences and quotients to the actual value. This was applied to frequently repeated features such as heart rate, oxygen saturation, temperature, systolic/diastolic/mean atrial blood pressure, respiration rate and serum glucose. 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 laboratory result and vital parameter to the actual date if more recent data were missing. This approach represents the medical perspective of decision making and laboratory results from blood samples are usually measured with varying frequency. Next, we made missingness explicit by introducing binary variables to indicate whether the values were carried forward. Additionally, we introduced numerical variables representing the up-to-dateness of the given value (0 for newly measured, 6 for measured 6 hours ago) so that machine learning models were able to learn the relevance of out-dated variables. Table 2 shows some derived features as described above for the mean atrial pressure (map) measured in the first 10 hours of a patient’s stay at the ICU. Empty cells indicating missingness of data which is explicitly tracked by

![Figure 1](https://example.com/figure1.png)

Figure 1. Data availability (black line) and sepsis proportion in subgroups of patients at specific hours post ICU admission (red line).

<table>
<thead>
<tr>
<th>Variables</th>
<th>Cumulative number of data points available (in %) with absolute number (n) of gaps/missing data between observations</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 0</td>
</tr>
<tr>
<td>Heart rate (hr)</td>
<td>1398740 (90%)</td>
</tr>
<tr>
<td>Oxygen saturation (o2sat)</td>
<td>1349202 (87%)</td>
</tr>
<tr>
<td>Body temperature (temp)</td>
<td>525111 (34%)</td>
</tr>
<tr>
<td>Systolic blood pressure (sbp)</td>
<td>1325945 (85%)</td>
</tr>
<tr>
<td>Mean atrial pressure (map)</td>
<td>1358496 (88%)</td>
</tr>
<tr>
<td>Diastolic blood pressure (dbp)</td>
<td>1065282 (69%)</td>
</tr>
<tr>
<td>Respiration rate (resp)</td>
<td>1313516 (85%)</td>
</tr>
<tr>
<td>End tidal carbon dioxide (et_co2)</td>
<td>57636 (4%)</td>
</tr>
</tbody>
</table>
the ‘miss’ variables. The robust variable generation was followed by the computation of assumed 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 and criteria were calculated:

- **ShockIndex** (hr/sbp)
- qSOFA (sbp and resp)
- SOFA and partial SOFA scores (respiration, renal function, platelets, liver function, sofa_renal, soфа_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:**
  - ShockIndex uses 25% and 75% quantiles of last 6h
  - qSOFA uses 25% and 75% quantiles of last 6h
  - SIRS uses 24 and partial scores uses 25%, 75% quantiles for temperature and 90% quantiles of the last 24h for heart rate and respiratory rate
  - NEWS uses 50% quantiles of respiratory rate, heart rate and systolic bp of the last 6h

The final size of the generated dataset was 1552210 × 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-based 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 (10% for non-sepsis, 200% for sepsis) and logloss as the stopping metric within the fitting processes. Furthermore, a stacked ensemble (SE) was built to further improve the predictability. Figure 2 illustrates our workflow.

### 5. Evaluation of Predictions & Results

We evaluated the threshold method by computing the normalized utility values $U_{\text{norm}}$ (see [8]) at each possible threshold in the training, validation and hold-out test set. The threshold was selected at the sepsis score with the maximal $U_{\text{norm}}$ 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 in our hold out set. 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 59.2% of all sepsis...
6. Conclusion

Time-specific meta-learners showed potential for a better understanding of the driving factors describing the pre-septic state depending on the hour after ICU admission. We have seen that frequently used clinical scores loses their ability for sepsis screening after the first day of ICU admission which demonstrates the need of better scores for routinely screening especially in ICU-acquired sepsis.

References


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