An Algorithm for Early Detection of Sepsis Using Traditional Statistical Regression Modeling

Roshan Pawar[1], Jeffrey Bone[1,2], J Mark Ansermino[1,2], Matthias Görges[1,2]

[1] The University of British Columbia, Vancouver, Canada
[2] Research Institute, BC Children's Hospital, Vancouver, Canada

Abstract

Sepsis is the final common pathway for many infections, whereby the body's immune response leads to organ failure, and eventually death. It is associated with high mortality rates and, if survived, significant morbidity. Early detection is imperative to improve outcomes. Yet, there is also a need to avoid a high false alarm rate. The aim of this study was to develop and evaluate a simple algorithm for early sepsis detection.

Significant missing data were encountered in the dataset, which were forward-filled or substituted with population means. Clinically relevant variable combinations were added along with transformation features including dichotomization, z-scores, first derivative, and changes from baseline. A logistic regression model was used to identify candidate features and build the overall risk score function for prediction.

The final candidate score had areas under the receiver operating characteristic curve of 0.747, 0.760, and 0.783 for the three test data sets. It had accuracies of 0.795, 0.889, 0.815, respectively, and an overall utility score for the full test set of 0.249 using a cutoff of 0.024.

Evaluation indicated significant potential for further optimization, including reduction of false-positive predictions. Adding features capturing change over time is expected to provide scope for further investigation.

1. Introduction

Artificial intelligence (AI) applications for clinical decision making and outcome prediction are demonstrating potential across the human lifespan [1], especially as access to integrated health care process and outcome data becomes easier. AI provides an opportunity to streamline clinical workflow and the promise of safer, more efficient, and more cost effective care [1], [2].

The intensive care unit (ICU) is a data-rich environment with a wide range of continuously monitored physiological variables that are responsive to clinical interventions over short time periods, and capture outcomes that are well-defined and generally quantifiable [3]. Thus, the ICU provides fertile ground for the development and evaluation of AI-based prediction models of individualized risks and outcomes.

Sepsis is the final common pathway of many infections, whereby the body's immune response leads to organ failure, and eventually death [4]. It is associated with high mortality rates; if survived, it results in significant morbidity [5]. In Canada, it was the 12th leading cause of death in 2011, with about 1 in 18 deaths involving sepsis [6].

Early detection and antibiotic treatment are critical for improving outcomes. Each hour of delayed treatment has been associated with a 4-8% increase in mortality [7]. However, there is also a need to avoid an overly high false alarm rate [8], [9], which places an unnecessary burden on healthcare resources and contributes to increasing costs. Significant barriers to clinically useful AI tools remain; these include model calibration, user trust, and data quality/heterogeneity [10].

The aim of this paper is to develop and evaluate a simple algorithm, using logistic regression, for early sepsis detection in adult ICU patients.

2. Methods

All analyses, other than the application of the competition-provided Python-based scoring tools, were performed using R 3.3.2 (R Foundation for Statistical Computing, Vienna, Austria).

2.1. Dataset

A labeled training set of time-series data from 40,336 patients admitted to an ICU was provided by The Early Prediction of Sepsis from Clinical Data - the PhysioNet Computing in Cardiology Challenge 2019 [11].

The reference set included 2,933 cases with a positive sepsis label (7.3%), with a median onset time of 29 (interquartile range [IQR] 7-73) hrs. The challenge includes a utility score to optimize against, which rewards classifiers that correctly predict sepsis between 12 hours before and 3 hours after clinical indication and penalizes

classifiers that fail to predict sepsis or predict sepsis more than 12 hours after. Hence, these requirements had to be taken into consideration to determine the optimal risk score threshold, instead of a sensitivity-/specificity-based threshold selection approach, for example.

2.3. Training data pre-processing

Significant missing values were encountered in the training dataset, which varied between 10-15% for some standard vital signs, to over 99% for some laboratory values. This was addressed by forward-filling of missing values, and, after calculating derivative variables, by substituting with either population means from non-sepsis patients, or means of the normal value ranges.

Minor data cleaning, such as the removal of impossible inspired oxygen concentration values [FiO_2], was necessary to avoid division-by-zero problems when normalizing values.

2.3. Feature generation

In addition to the raw values provided, three areas for feature generation were explored: 1) combinations of existing variables that were deemed clinically relevant, 2) transformations of candidate variables, and 3) evaluation of time-based changes.

2.3.1 Derived variables of potential clinical interest

Derived cardiac-based clinical features included: pulse pressure [PP] (systolic blood pressure [SBP]-diastolic blood pressure), estimated cardiac output [CO] (pulse pressure times heart rate [HR]), shock index (HR/SBP) and modified shock index [mSI] (HR/mean arterial blood pressure [MAP]) [12], cardiac output variation (PP/mean arterial pressure [MAP]) [13], and temperature-adjusted HR and respiratory rate [RR].

For respiratory-based features, we transformed oxygen saturation based on the concept of virtual shunt (VS) [14], the difference between predicted (using temperature and HR) and measured respiratory rate [Resp], Carrico index (arterial oxygen partial pressure [O_2Sat or SaO_2] to FiO_2 ratio), oxygen delivery (combining hemoglobin, O_2Sat, and CO or MAP).

For laboratory test-derived variables, we calculated the number of laboratory measurements available at a given time, assuming that, as patient severity increased, additional investigations were performed for monitoring and therapy adjustments. Further, we derived urea/creatinine ratios, bicarbonate [HCO_3^-]/lactate ratios, calculated the anion gap (assuming normal sodium concentration), and computed linear combinations of urea and creatinine, HCO_3^- and lactate, chloride and pH as well as O_2Sat- and MAP-adjusted hemoglobin.

2.3.2 Variable transformations

Firstly, we dichotomized all candidate features, whereby each observation was compared to reference normal values in Medscape (WebMD, New York, NY). Missing data were considered normal. The missingness status itself was also captured as a candidate variable.

Next, we obtained z-scores using the mean and standard deviation of observations from patients who never had a positive sepsis label. We also included an absolute version of the z-score as a feature, as it might simply be the deviation from normal (not its direction) that needs to be assessed.

Further, we created two different "penalty scores": the absolute difference from the non-sepsis patients' mean observations, set to zero if it fell within the normal range for the given variable; and using the absolute z-score, also set to zero if it fell within the normal range.

Finally, we added the count of abnormal variables per organ system, specifically cardiac (HR, MAP, SBP, and troponin), respiratory (O_2Sat, end-tidal carbon dioxide concentration, SaO_2, HCO_3^-, lactate and the Carrico index), coagulation (fibrinogen, platelets, and partial thromboplastin time), and an infection category (white blood cell count and temperature).

2.3.3 Time-based features

Two simple change indices were created to investigate temporal relationships: 1) the first derivative for all values observed, and 2) the change from baseline observation. For the latter, we used the first observations after ICU admission as the reference.

2.4. Feature selection

For each of the 34 provided variables, and our derived features, we built logistic regression models using all variable transformations, as well as the values themselves. From these candidate variables, we selected those with a highly significant p-value (<0.01) for consideration in the sepsis identification algorithm. In addition, we included the number of laboratory values, and patient age and sex in the final regression model.

2.5. Model performance evaluation

A logistic regression model was trained and the initial risk score threshold identified using the Youden cut-off from the receiver operating characteristics (ROC) curve. Confidence intervals (CIs) for the area under the ROC curve (AUROC) were obtained through bootstrapping, but no other model cross-validation was performed. Model coefficients were then implemented into a risk scoring function. Additionally, model performance was

evaluated using the competition-provided scoring cost function. After obtaining the utility score based on the Youden threshold, additional optimization was performed by varying the probability cutoff over 70% (p=0.25-0.95) of the observed probability range, in order to gain the highest utility score for a given model.

3. Results

Our initial competition score entry had an AUROC of 0.704, an area under the precision-recall curve (AUPRC) of 0.065, an accuracy of 0.781, and an official utility score of 0.200, when using a probability cutoff of 0.0155. This model used only commonly measured vital signs (HR, O_2Sat, Temp, SBP, MAP, DBP, Resp), derived indices CO, PP, and mSI, and age and sex, with the variable itself, its first derivative, and absolute z-score, as features. Optimizing the cutoff resulted in a maximum utility score of 0.238 when using a cutoff of 0.022.

When evaluating different variations of the z-score as the sole predictors, the absolute z-scores resulted in marginally superior model performance with an AUROC of 0.720, when compared to 0.717 for the standard z-score, 0.680 for the penalty z-score, and 0.676 for the absolute penalty z-score. The absolute z-score was used for all future evaluations.

Evaluating only commonly measured variables and retaining only statistically significant variables in the logistic regression (age, sex, O_2Sat, Temp, SBP, Resp, CO, SI, VS, urea nitrogen, creatinine, glucose, potassium, hematocrit, hemoglobin, white blood cell counts, platelets, and urea/creatinine ratio), yielded an AUROC of 0.783. However, if missing values were substituted with the mean from non-sepsis patients, to allow the model to be applied to all samples, the AUROC decreased to 0.728, and if replacing them with the mean value of the normal range, the AUROC decreased to 0.745. This is likely due to the imputation method being uninformative and the case mix of patients with missing data being very diverse.

A model for each feature type, using only statistically significant variables as indicated in their respective logistic regression, along with age, sex, and number of laboratory values measured, had AUROCs of 0.730 for variable abnormality, 0.720 for variable absolute z-scores, 0.674 for the first derivative of the variable, 0.576 for missingness of the commonly measured vital signs, and 0.678 for deviations from baseline values.

When combing all of these variables, except for the baseline deviation variables, in a single new model, it had an AUROC of 0.790 (95%CI 0.787-0.792) (see Fig. 1) during model building; when applying the `get_sepsis_score.R` function to the training data, it resulted in an AUROC of 0.774, an AUPRC of 0.078, an accuracy of 0.844, and a utility score of 0.301 for a probability cutoff of 0.024.

The **official utility score** for the full test set was 0.249, with scores of 0.296, 0.268, and 0.007 for test sets A, B, and C, respectively. Official AUROCs were 0.747, 0.760, and 0.783; AUPRCs were 0.072, 0.067, and 0.088; and accuracies were 0.795, 0.889, and 0.815 respectively.

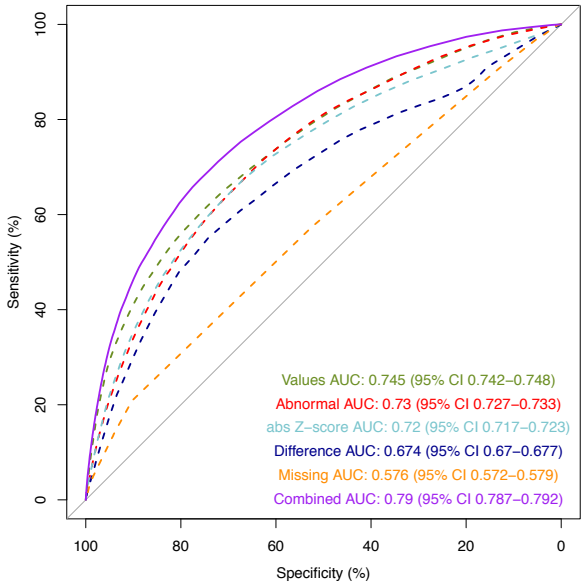

Figure 1. Receiver operating characteristics curve of model performance (final attempt) on the training dataset. The AUC and 95% confidence interval (CI) for each approach are indicated in the bottom corner. '*Abnormal*' indicates the use of variable normality, 'abs *Z-score*' the use of absolute variable z-scores, '*Difference*' the use of the 1st derivatives, '*Missing*' the use of vital signs missingness, and '*Combined*' for a combination of all previously mentioned features.

4. Discussion

The goal of this model building study was to create a sepsis prediction model, which used only simple features and explainable machine learning approaches: specifically variable transformations and logistic regression. The lack of model performance gain when using missing vital signs features was surprising as we had assumed (incorrectly) that additional monitoring would be correlated with higher acuity, and thus higher sepsis probability.

Despite the introduction of additional new features since our initial attempt, including clinically-relevant variable transformations and variable missingness features, we failed to make significant improvements in model performance. Our final performance, with AUROCs between 0.747 and 0.783, was considerably lower than that reported by Nemati *et al.*, with AUROCs of 0.83-0.85 [15].

4.1. Future work

Scope for further investigation includes creation of additional features, particularly exploiting slower (longer time-scale) changes, the laboratory value missingness patterns, and the creation of additional interaction variables between physiological vital signs and laboratory values, with a particular emphasis on identifying organ system dysfunction.

Next, more sophisticated approaches to address data missingness, such as multivariate imputation by chained equations (MICE), may yield better performance [16], [17], as the substituted values are likely closer to specific cases than the overall sample.

In addition, the selection of variable thresholds to determine abnormality could be tailored to be more specific to the ICU setting, instead of using general population reference values. This is important as some of the interventions performed, such as mechanical ventilation, will mean some measured values will necessarily be different to those observed in otherwise healthy subjects.

Finally, the use of more sophisticated machine learning approaches [18] might yield additional performance gains; these approaches may need to be explainable to maintain clinician trust in the derived predictions.

4.2. Conclusion

Our candidate score showed moderate performance with a AUROC between 0.747 and 0.783 against the test data, and a utility score of 0.249 for the full test set; it received a rank of 35/78 entries. Evaluation indicated significant potential for further optimization, including reduction of false-positives. Additional change-over-time features are expected to provide valuable scope for further investigation.

Acknowledgments

The authors wish to thank Nicholas West for editorial assistance. This work was supported, in part, by a Natural Sciences and Engineering Research Council of Canada grant (RGPIN-2018-05121).

References

[1] E. J. Topol, "High-performance medicine: the convergence of human and artificial intelligence.," *Nat. Med.*, vol. 25, no. 1, pp. 44–56, Jan. 2019.

[2] C. D. Naylor, "On the prospects for a (deep) learning health care system," *JAMA*, vol. 320, no. 11, p. 1099, Sep. 2018.

[3] A. E. W. Johnson, M. M. Ghassemi, S. Nemati, K. E. Niehaus, D. A. Clifton, and G. D. Clifford, "Machine learning and decision support in critical care.," *Proc. IEEE. Inst. Electr. Electron. Eng.*, vol. 104, no. 2, pp. 444–466, Feb. 2016.

[4] M. Singer *et al.*, "The third international consensus definitions for sepsis and septic shock (Sepsis-3).," *JAMA*, vol. 315, no. 8, pp. 801–10, Feb. 2016.

[5] S. L. Weiss *et al.*, "Global epidemiology of pediatric severe sepsis: the sepsis prevalence, outcomes, and therapies study.," *Am. J. Respir. Crit. Care Med.*, vol. 191, no. 10, pp. 1147–57, May 2015.

[6] T. Navaneelan, S. Alam, P. A. Peters, and O. Phillips, "Health at a glance: deaths involving sepsis in Canada," Ottawa, ON, 2016.

[7] A. Kumar *et al.*, "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock.," *Crit. Care Med.*, vol. 34, no. 6, pp. 1589–96, Jun. 2006.

[8] M. Borowski, M. Görges, R. Fried, O. Such, C. Wrede, and M. Imhoff, "Medical device alarms," *Biomed. Tech. (Berl).*, vol. 56, no. 2, pp. 73–83, Mar. 2011.

[9] L. Varpio, C. Kuziemsky, C. MacDonald, and W. J. King, "The helpful or hindering effects of in-hospital patient monitor alarms on nurses: a qualitative analysis.," *Comput. Inform. Nurs.*, vol. 30, no. 4, pp. 210–7, Apr. 2012.

[10] N. D. Shah, E. W. Steyerberg, and D. M. Kent, "Big data and predictive analytics," *JAMA*, vol. 320, no. 1, p. 27, Jul. 2018.

[11] M. A. Reyna *et al.*, "Early prediction of sepsis from clinical data: the PhysioNet/Computing in Cardiology challenge 2019," *Crit. Care Med.*, p. In press, 2019.

[12] Y.-C. Liu *et al.*, "Modified shock index and mortality rate of emergency patients.," *World J. Emerg. Med.*, vol. 3, no. 2, pp. 114–7, 2012.

[13] A. Tantot *et al.*, "Evaluation of cardiac output variations with the peripheral pulse pressure to mean arterial pressure ratio.," *J. Clin. Monit. Comput.*, vol. 33, no. 4, pp. 581–587, Aug. 2019.

[14] G. Zhou, W. Karlen, R. Brant, M. Wiens, N. Kissoon, and J. M. Ansermino, "A transformation of oxygen saturation (the saturation virtual shunt) to improve clinical prediction model calibration and interpretation.," *Pediatr. Res.*, Aug. 2019.

[15] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman, "An interpretable machine learning model for accurate prediction of sepsis in the ICU.," *Crit. Care Med.*, vol. 46, no. 4, pp. 547–553, Apr. 2018.

[16] S. van Buuren and K. Groothuis-Oudshoorn, "mice : multivariate imputation by chained equations in R," *J. Stat. Softw.*, vol. 45, no. 3, pp. 1–67, 2011.

[17] M. G. Kenward and J. Carpenter, "Multiple imputation: current perspectives.," *Stat. Methods Med. Res.*, vol. 16, no. 3, pp. 199–218, Jun. 2007.

[18] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainabilty of artificial intelligence in medicine," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 9, no. 4, p. e1312, Apr. 2019.

Address for correspondence:
Matthias Görges, V3-324, BC Children's Hospital Research Institute, 950 West 28th Ave, Vancouver, BC, V5Z 4H4, Canada; mgorges@bcchr.ca