Jonne Kurokallio & Vesa Vahermaa

Early Prediction of Sepsis from Clinical Data Using Generative Longitudinal Modelling

Abstract

Sepsis is one of the leading causes of death in intensive care unit (ICU) patients. Being able to react quickly to sepsis onset or successfully predict sepsis could significantly lower the mortality rates of ICU patients. In this work, we examine the viability of using various machine learning methods to identify time points when sepsis is imminent or has recently started. The models use the vital signs, laboratory values, and demographic information of ICU patients for classification. Our results show that these models can be used to predict and identify sepsis at above-chance levels, although they do not reach or exceed the results obtained by other similar models in previously published literature.

1. Introduction

Sepsis is among the leading causes of death in patients admitted to intensive care. While the mortality rate for patients who undergo sepsis remains high, they have declined significantly over time (Stevenson, Rubenstein, Radin, Wiener & Walkey, 2014). Being able to administer care quickly after sepsis occurs is vital, as a shorter reaction time to administering sepsis care is associated with lower mortality (Seymour et al., 2017). For this reason, novel methods that aim to predict sepsis before it occurs may eventually give doctors an invaluable tool for saving patients at a higher rate.

Recent research of machine learning in the context of sepsis provides some hope: for example *InSight*, a neural net classifier using multivariate combinations of easily obtained patient data was able to identify sepsis at a higher rate (0.88 AUROC at time of sepsis onset) than other clinical prediction tools such as Sequential Organ Failure Assessment (SOFA) and Systemic Inflammatory Response Syndrome (SIRS) prior to sepsis onset (Desautels et al., 2016). A more recent mixed-ward study validates *InSight's* performance as outperforming existing sepsis scoring systems in identifying and predicting sepsis, severe sepsis, and septic shock (Mao et al., 2018).

Other more recent approaches have also achieved similarly promising results: a model using a modified regularized Weibull-Cox analysis achieved a similarly high rate of sepsis identification (0.85 AUROC), this time 4 hours before sepsis onset (Nemati et al., 2018).

Consensus is yet to be reached on which features and models yield best possible prediction results. The recent PhysioNet Computing in Cardiology Challenge 2019 invited researchers to design and implement open-source algorithms that identify and predict sepsis based on vital signs and other available clinical data (Reyna et al., 2020). This report and the accompanying model is based on the PhysioNet challenge.

2. Materials and Methods

2.1. Data

The data set for the competition was collected from ICU patients in multiple hospital systems, but for this project only the A and B datasets were available. The data consist of three parts: Vital signs (8 features), Laboratory values (26 features) and demographics (7 features). Altogether there are 40336 patients and the previously mentioned values are taken at an hourly basis. There exist vast differences between patients' time in the ICU care. In other words some patients have hundreds of rows of data, while some may have only ten.

Undoubtedly the largest difficulty with this data set is the huge number of missing values, denoted as NaN. As can be seen from ***fig. 1*** 27 features are missing more than 80% of their values, some even more than 95%. Missing values greatly affect the prediction and training process, and as they are so frequent in some features, it is hard to simulate the data distributions in those features. Therefore for comparison the final models will be trained with both all data and with data, where the features with more than 90% of values missing are reduced.

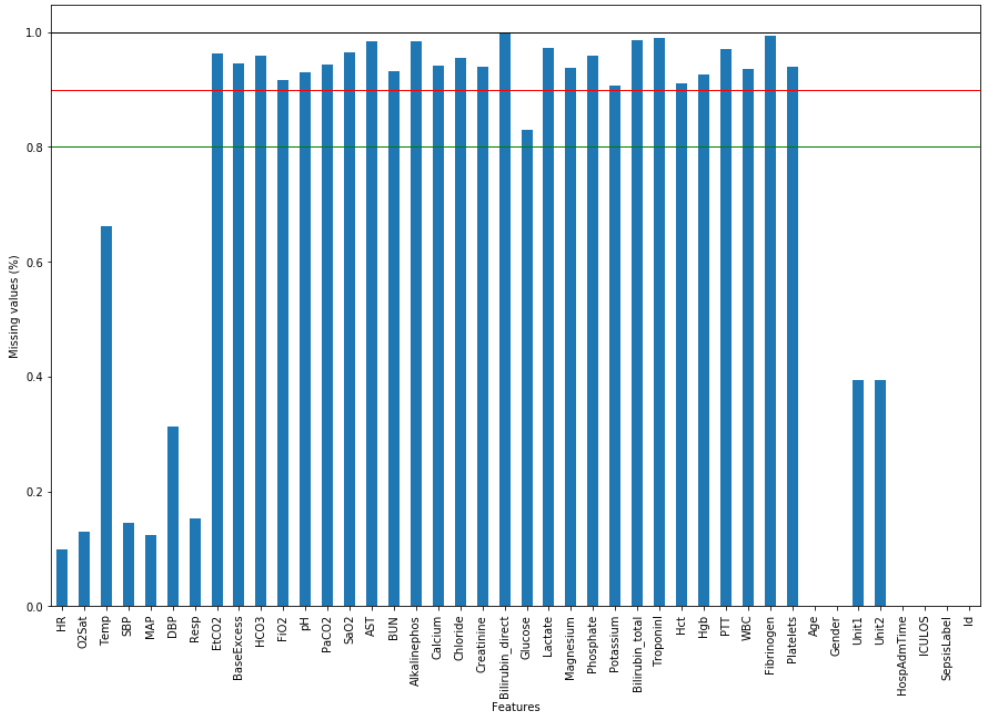


Fig. 1. The percentage of missing values per feature

Another attribute of the dataset worth paying attention to was the distribution of sepsis labels for each patient. In the data set the amount of patients that eventually will get sepsis was only 7.27%. This in mind the data was divided into training and testing sets in such a way that 20% of sepsis patients were considered in the test set. This had a minor increase in the utility score in comparison to a random train/test distribution of the data. The initial sepsis label was also separated into two different covariates: time to/from sepsis and sepsis label. The main reason for this was to be able to use the given utility function better. The time to/from sepsis holds the values of negative hours before sepsis, and after sepsis, the positive hours from the first signs of sepsis. The sepsis label is a binary label and tells if the patient is at some point going to suffer from sepsis.

When inspecting the data and contrasting the distributions of various values of individuals with identified sepsis and those with no identified sepsis, clear differences could be identified. Some of the largest differences can be seen in the heart rate, body temperature, hemoglobin, platelet, and creatinine levels, as seen in appendices A to E. These graphs include an equal number of individual measurement points taken from rows with SepsisLabel = 0 and SepsisLabel = 1. The differences are less pronounced when comparing pre-sepsis measurement points (up to 12 hours before sepsis) to other measurement points with no sepsis.

In the python script the data is concatenated in one large dataframe one below another and then normalized. As mentioned the train/test division is done by 80% / 20%.

2.2. Utility function and scoring method

The scoring method for every algorithm attending to the competition was determined beforehand by the organizers. A utility function was given to participants so they could evaluate the binary classification performance of their models. The utility function takes account both cases of positive true/false and negative true/false predictions and rewards or penalizes depending on the time of sepsis prediction. The utility score per sepsis prediction can be seen in *fig 2.*.

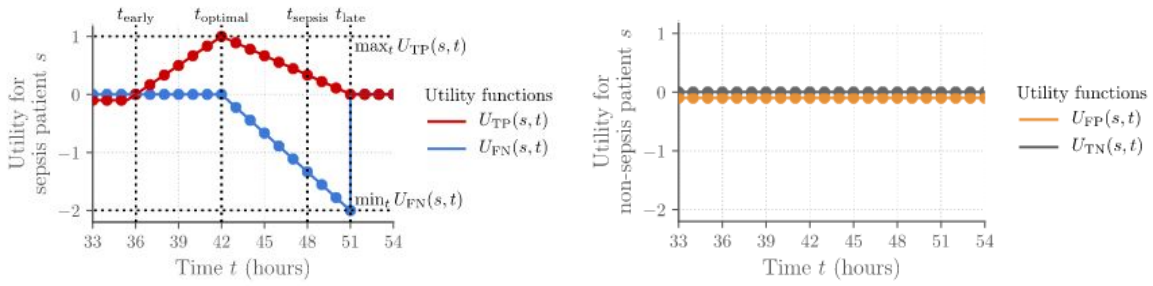


Figure 2. The utility score based on the binary classification and time of prediction

The utility function puts the weight heavily on the true positive and false negative sepsis prediction, as those predictions are the most important when it comes to the survival of the patient. For early prediction of the sepsis the score is close to zero but increases linearly when time is getting closer to the sweet spot. In the data set the sepsis label set as 1 meant, that the sepsis is going to take place within the next 6 hours. The time of sepsis is the first clinical suspicion of infection or when the SOFA score of the patient drops within 24-hour period. Keeping that in mind it makes sense to give the most points to right positive prediction at that time stamp. Late true positive predictions are not penalized, because the organ damage does not take place exactly at the time of sepsis and there might still be time to save the patient. False negatives are penalized heavily after the optimal sepsis prediction time, because if this prediction is followed blindly it will eventually kill the patient.

More comparison to different scoring method in final report.

2.3. Trained Models

2.3.1 Generative models

One of the main tasks in this project was to create multiple generative models to generate joint probability distribution for the data. Conditional probability *P(X|Y=y)* can be calculated from this joint probability, where *X* is the observable covariates and *Y* is the target. In other words the aim is to model the features and covariates.

It is assumed that the data follows normal distribution with mean vector μ and standard deviation matrix *D*. Our generative models follow notation where $Y = [y_1{}^T, ..., y_N{}^T]^T$ is the matrix of modelled outcomes and $X = [x_1{}^T, ..., x_N{}^T]^T$ is the matrix of covariates. N is the number of rows in the data set, i.e. data points. As mentioned the Y is assumed to be normally distributed as $y_i \sim N(\mu(x), D)$, where $\mu(x)$ is the function that maps relationship between covariates and outcome expectations in each row of the data set.

In the first model the outcome expectations are assumed to be linearly dependent on the covariates.The mean vector is formed by $\mu(x) = Bx$, where B is the matrix of linear regression coefficients. The trainable parameters in this model are B and standard deviation matrix *D*, which is a diagonal matrix ($D = diag(\sigma^2{}_1, .., \sigma^2{}_j)$). This model assumes that the mean vector of the data distribution can be linearly drawn from the covariates, which is not the case in many real life scenarios. For example this model is unable to detect, if old people and young people are more presented in sepsis patients than the middle aged. In such case the relationship should resemble more of a parabola for example. (Maybe draw here picture to demonstrate)

The second model uses normal multilayer perceptron (MLP) to apply non-linear relationship between covariates and outcome expectations ($\mu(x) = f_\theta(x)$, where f is fully connected MLP and θ are the trainable parameters). Multiple numbers of hidden layers and layer sizes were tested in the MLP. The initial structure was two hidden layers with 14 and 28 nodes respectively followed by Relu- activation function. The non-linearity enables the modelled outcome distributions to fall on a curve instead of a straight line imitating real life data more than the previous model. (Draw picture to demonstrate) The covariance matrix is assumed to be diagonal in this model as well.

In the third model the standard deviation matrix is no longer considered to be diagonal. The noise is considered un-independent which means that the covariance matrix is formed as:
$\Sigma_{ij} = \text{cov}(Y_i, Y_j)$
Similarly to the previous model, the MLP allows the generated mean vector to have a curvy quality ($\mu(x) = f_\theta(x)$). The residuals however do not have equal variance and will in this case affect each other. (Draw picture to demonstrate)

The math for each model is almost similar with minor changes. The probability distribution function for the normal distribution $N(\mu, \sigma^2)$ is

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} * e^{\frac{-(x-\mu)^2}{2\sigma^2}},$$

from which we can create the log-likelihood function to estimate the parameters the following way:

$$ln\, L(\mu, \sigma^2|x_i) = \sum_{i=1}^{n} ln\, f(x|\mu, \sigma^2) = -\tfrac{n}{2} ln\,(2\pi) - \tfrac{n}{2} ln\, \sigma^2 - \tfrac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 .$$

The log-likelihood function can be written in matrix notation for our models (select $\mu(x)$ and *D* according to the model) as:

$$ln\, L(\mu(X), D|X) = \sum_{i=1}^{n} ln\, f(X|\mu, D) = -\tfrac{n}{2} ln\,(2\pi) - \tfrac{n}{2} ln\, D - \frac{(Y-\mu(X))^T (Y-\mu(X))}{2D}$$

This log-likelihood is used as loss function in the training loop.

The models and parameters are tested more thoroughly later in the project.

2.3.2 Long Short Term Memory (LSTM) model

Outside of the project scope recurrent neural network (discriminative model) to add in the comparison. This will be done last when the previous three generative models are covered.

2.3.3 Classification model

For the final classification task a “common sense”- model was received as a template. This will be replaced with a better one once the three generative models have been scripted.

3. Results

The models are going to be tested with multiple parameter values (especially with the MLP) and data set modifications. Below is a short summary how the testing is probably going to be carried out.

Model	DataSet	Parameters	Utility score
Linear model with heterogeneous noise.	Full randomly divided training set --- Full 80/20 by sepsis patient divided dataset --- Dataset where features with >90% missing values reduced	B and D	
Non-linear model	Full randomly divided	N_Layers =	

with heterogeneous noise.	training set --- Full 80/20 by sepsis patient divided dataset --- Dataset where features with >90% missing values reduced	(3),(5),(10) Nodes per layer = Try few values	
Non-linear model with noise correlated across the outcomes	Full randomly divided training set --- Full 80/20 by sepsis patient divided dataset --- Dataset where features with >90% missing values reduced	N_Layers = (3),(5),(10) Nodes per layer = Try few values	
Long Sort Term Memory	Dataset where features with >90% missing values reduced		

4. Discussion & Conclusions

While our models failed to outperform the PhysioNet challenge top scorers, they performed clearly above chance levels, validating the concept that machine learning methods can successfully use patients' vital signs in order to identify imminent or recent sepsis.

One of the most significant difficulties in using the data was the lack of laboratory measurements. It is understandable that laboratory test values are less common than vital sign values, as collecting laboratory test values is costly, can cause patient discomfort and risk potential infection; furthermore, more frequent testing may not even yield significantly better patient outcomes (Ezzie, Aberegg & O'Brien, 2007; Kotecha, Shapiro, Cardasis & Narayanswami, 2017).

Note: More thorough discussion & conclusion will be included in the final report, with proper results.

5. References

Desautels, T., Calvert, J., Hoffman, J., Jay, M., Kerem, Y., & Shieh, L. et al. (2016). Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach. *JMIR Medical Informatics*, *4*(3), e28. doi: 10.2196/medinform.5909

Ezzie, M., Aberegg, S., & O'Brien, J. (2007). Laboratory Testing in the Intensive Care Unit. *Critical Care Clinics*, *23*(3), 435-465. doi: 10.1016/j.ccc.2007.07.005

Kotecha, N., Shapiro, J., Cardasis, J., & Narayanswami, G. (2017). Reducing Unnecessary Laboratory Testing in the Medical ICU. *The American Journal Of Medicine*, *130*(6), 648-651. doi: 10.1016/j.amjmed.2017.02.014

Mao, Q., Jay, M., Hoffman, J., Calvert, J., Barton, C., & Shimabukuro, D. et al. (2018). Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. *BMJ Open*, *8*(1), e017833. doi: 10.1136/bmjopen-2017-017833

Nemati, S., Holder, A., Razmi, F., Stanley, M., Clifford, G., & Buchman, T. (2018). An Interpretable Machine Learning Model for Accurate Prediction of Sepsis in the ICU. *Critical Care Medicine*, *46*(4), 547-553. doi: 10.1097/ccm.0000000000002936

Reyna, M., Josef, C., Jeter, R., Shashikumar, S., Westover, M., & Nemati, S. et al. (2020). Early Prediction of Sepsis From Clinical Data. *Critical Care Medicine*, *48*(2), 210-217. doi: 10.1097/ccm.0000000000004145

Seymour, C., Gesten, F., Prescott, H., Friedrich, M., Iwashyna, T., & Phillips, G. et al. (2017). Time to Treatment and Mortality during Mandated Emergency Care for Sepsis. *New England Journal Of Medicine*, *376*(23), 2235-2244. doi: 10.1056/nejmoa1703058

Stevenson, E., Rubenstein, A., Radin, G., Wiener, R., & Walkey, A. (2014). Two Decades of Mortality Trends Among Patients With Severe Sepsis. *Critical Care Medicine*, *42*(3), 625-631. doi: 10.1097/ccm.0000000000000026

---------- NOT USED (at least not yet) -------
Islam, M. M., Nasrin, T., Walther, B. A., Wu, C. C., Yang, H. C., & Li, Y. C. (2019). Prediction of sepsis patients using machine learning approach: a meta-analysis. Computer methods and programs in biomedicine, 170, 1-9.

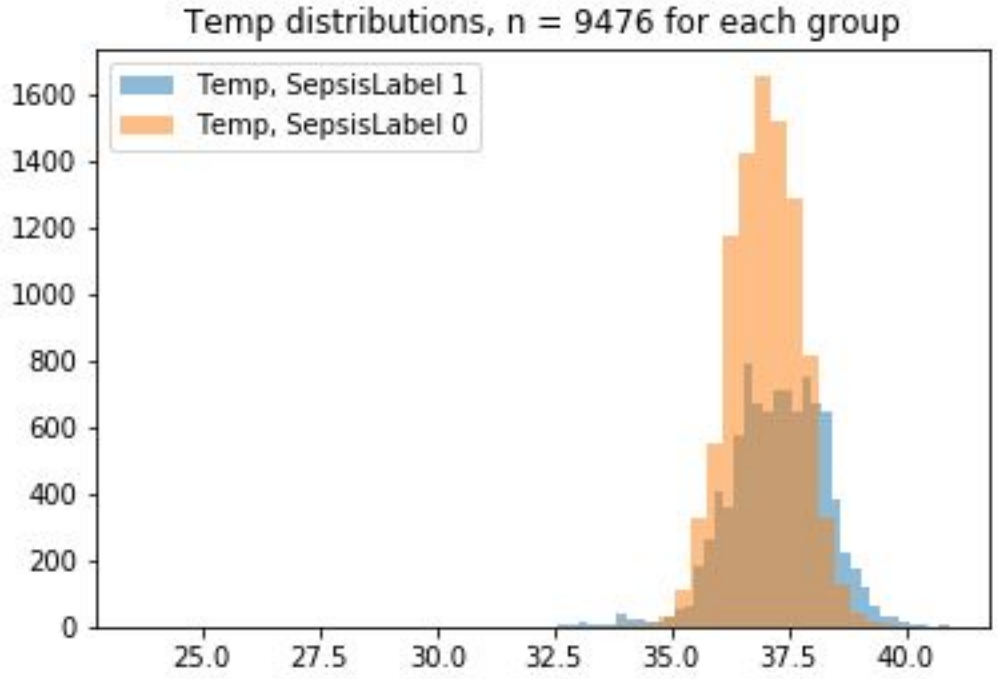
Lukaszewski, R. A., Yates, A. M., Jackson, M. C., Swingler, K., Scherer, J. M., Simpson, A. J., ... & Pearce, M. J. (2008). Presymptomatic prediction of sepsis in intensive care unit patients. Clin. Vaccine Immunol., 15(7), 1089-1094.

[Luo et al., 2016] Luo, J., Wu, M., Gopukumar, D., and Zhao, Y. (2016). Big data application in biomedical research and health care: A literature review. Biomedical Informatics Insights, 8:BII.S31559

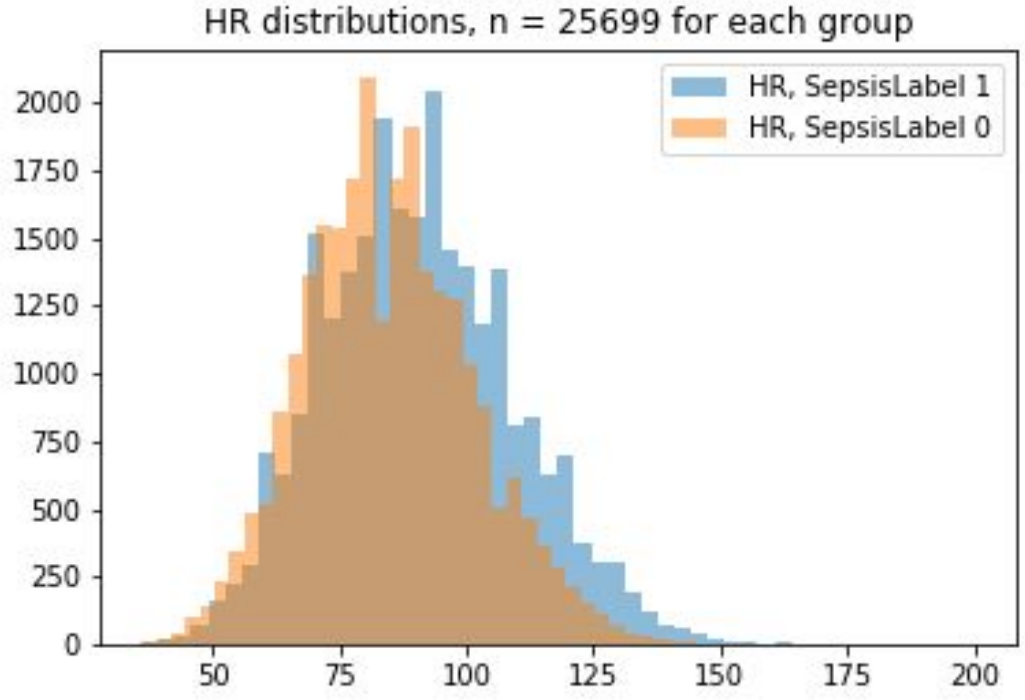
[Obermeyer and Emanuel, 2016] Obermeyer, Z. and Emanuel, E. J. (2016). Predicting the future big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13):1216–1219.

[Hoi et al., 2006] Hoi, S. C. H., Jin, R., Zhu, J., and Lyu, M. R. (2006). Batch mode active learning and its application to medical image classification. Proceedings of the 23rd international conference on Machine learning - ICML ’06

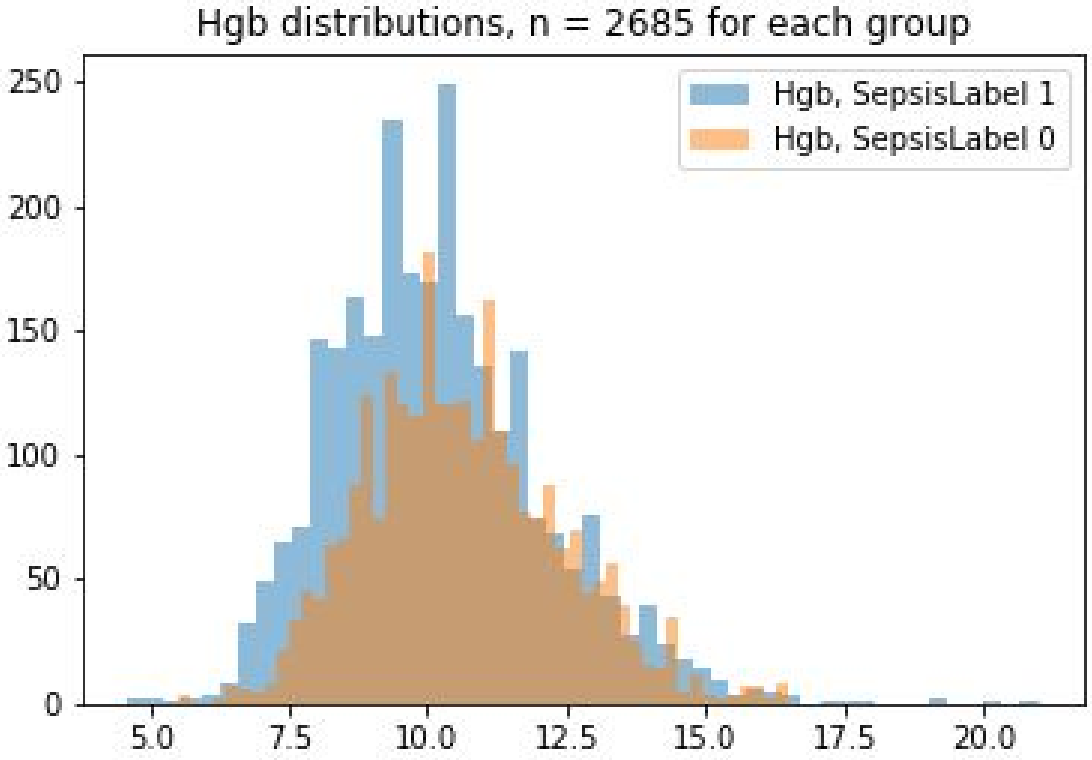
6. Appendices



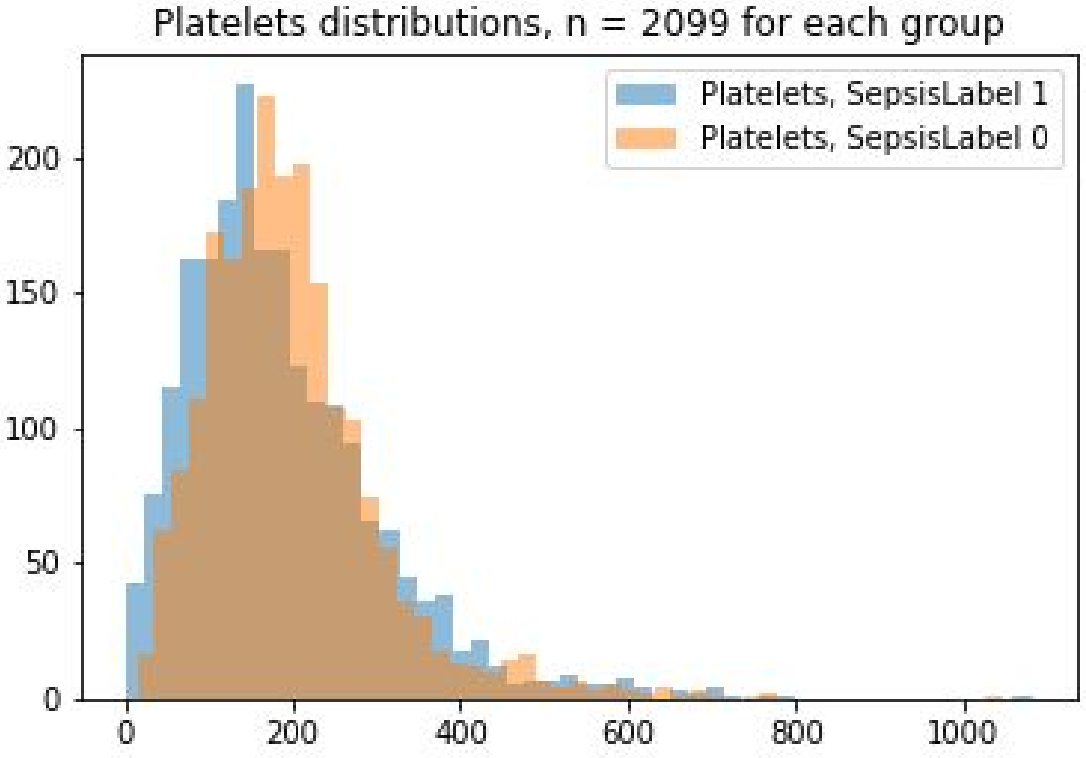
Appendix A. Body temperatures in pre/no-sepsis state and state of sepsis.



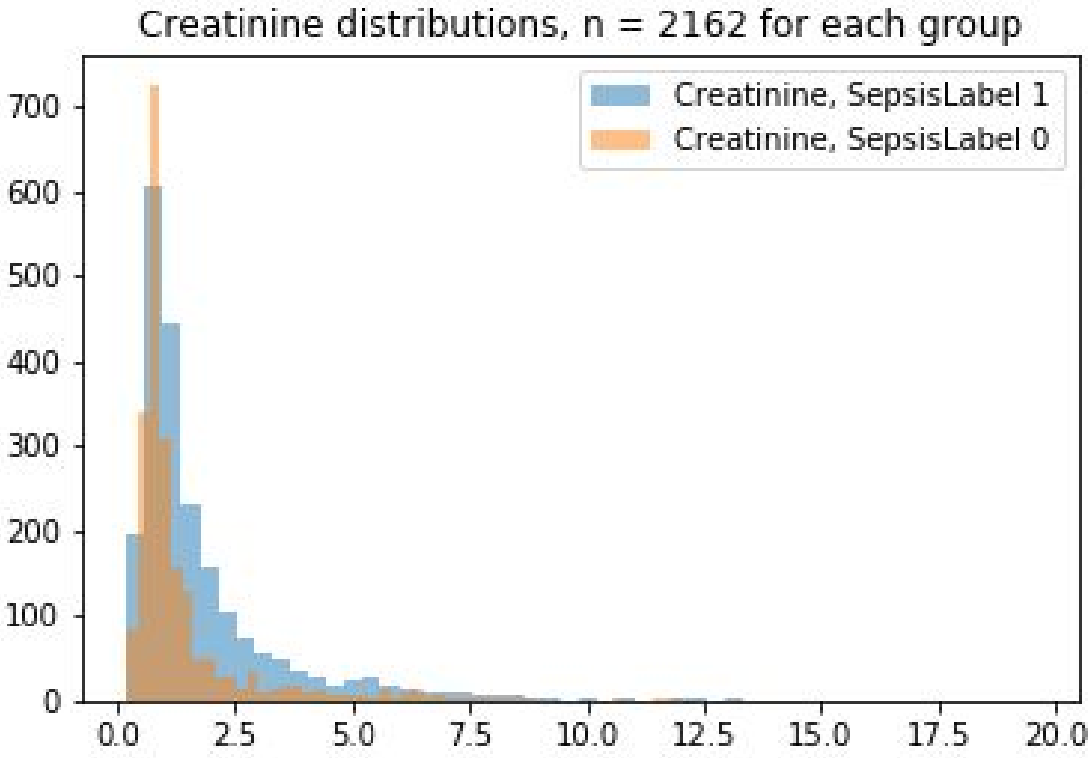
Appendix B. Heart rates in pre/no-sepsis state and state of sepsis.



Appendix C. Hemoglobin levels in pre/no-sepsis state and state of sepsis.



Appendix D. Platelet levels in pre/no-sepsis state and state of sepsis.



Appendix E. Creatinine levels in pre/no-sepsis state and state of sepsis.

7. Division of labor

The course work was done in a group of two, participants Jonne Kurokallio and Vesa Vahermaa. Jonne's main responsibility was for developing the learning models that achieved the actual results, while Vesa did data analysis and visualization as well as literature review. Both participants contributed to writing of the report, with Vesa writing all of chapter 1 and most of 4, and Jonne writing most of 2 and all of 3.