Early Sepsis Detection Driven By Personalized Patient Similarity Models

by

Jordyn Walton

A research paper presented to the
University of Waterloo
In partial fulfillment of the research requirement for the degree of
Masters of Mathematics
in
Statistics

Waterloo, Ontario, Canada, 2020

Abstract

Sepsis is a common and life-threatening condition that's difficult to diagnose. Therapies which treat sepsis early are able to improve health outcomes. However, early recognition of sepsis is a difficult task. In this paper, we deploy patient similarity models to detect sepsis 6 hours early using electronic health records and optimize for the number of similar patients. We find the addition of patient similarity to statistical learning models achieves promising(?) results for early sepsis detection, outperforming (?) population level models. Our testing of similarity models also validated previous results that showed predictive performance would suffer if the number of similar patients selected was too small.

Acknowledgements

I wanted to thank my supervisor, Dr. Joel Dubin for his guidance and dedication to my essay in these crazy times.

1 Introduction

Sepsis is a major public health concern and a common, life-threatening condition that occurs when the body's response to infection causes tissue damage, organ failure or death [12]. Each year, over 18 million people develop severe sepsis worldwide and it is associated with a 30-50% mortality rate [3]. Therapies which are able to manage severe sepsis and septic shock early and aggressively are able to improve health outcomes. Consequently, therapies which are delayed due to misidentification of sepsis are associated with worse health outcomes. Each hour of delayed treatment is associated with roughly an 4-8% increase in mortality [9][10]. The risk is increased in patients suffering from septic shock where an hourly delay can have as high of an increase of 9.9% in mortality [7]. However early identification of sepsis is a difficult task.

Sepsis can be difficult to diagnose. The signs and symptoms of sepsis are considered to be highly variable. They can differ according to several factors such as the patient, the pathogen resulting in the condition and how the condition evolves [6]. Although new clinical criteria that aid diagnosis has been developed, there is still an unmet need for early detection [9][10].

In response to this unmet need, the early prediction of sepsis from clinical data was chosen as the Computing in Cardiology Challenge of 2019. The challenge had the aim to facilitate the development of new automated open-source algorithms and to directly compare them on an a common ICU dataset. Participants were assigned the task of using detailed physiological data in the form of electronic health records involving hourly heart rate, blood pressure and other vital signs and lab values to detect sepsis 6 hours before clinical prediction of sepsis. As a result of the competition, the hourly physiological data from ICUs in two hospital systems was made freely available.

A multitude of algorithms were implemented and submitted to the conference, however none involved the use of patient similarity. Models driven by patient similarity will identify and train on a set of past patents which are similar to the case of interest, or the index patient, whose outcome we are aiming to predict. Personalized data prediction in clinical medicine is still a developing field, however patient similarity approaches have shown to be successful with prediction of patient outcomes and has improved upon performance of population-based models[5][11]. As well, analogous approaches have successfully been applied in other domains such as collaborative filtering for personalized product recommendation and future career trajectory for Major League Baseball players in sports analytics[5].

Patient similarity could be a favourable addition to statistical learning approaches for early sepsis detection models by leading to prediction which is more personalized and precise.

1.1 Outline

Our goal was to test whether the addition of patient similarity would improve performance of sepsis prediction models.

The objectives were 1) to build prediction models of sepsis based only on the patient's first 6 hours in the ICU, identifying the risk of a patient developing sepsis 6 hours later and making a positive or negative prediction and 2) to integrate patient similarity into the models and assess whether the addition would improve model performance. Special consideration would be given to incorporating longitudinal data into the prediction model as well as nuanced problems in the available physiological data, such as missing data and class imbalance.

1.2 Patient Similarity

The addition of patient similarity has previously improved predictive performance for models driven by electronic medical records. [5] The idea behind patient similarity models, is being able to optimize the predictive power of patients in the model. The underlying hypothesis of patient similarity analytics is that the amount of predictive power contributed by a patient should be directly proportional to the degree of similarity to the index patient citeJoelSimilarity. The converse of the hypothesis is that including data from dissimilar patients will have a negative effect, degrading predictive performance. Therefore, allowing a model to selectively train on similar patients allows the model to maximize predictive performance.

Nonetheless, an important consideration for patient similarity is the threshold chosen for number of similar patients. If too few are selected, the predictive model will begin to suffer from small-sample effects, limiting any improvement in performance and offsetting the positive effect of a more personalized model.

Special consideration can be taken to consider how to define patient similarity. Several approaches for defining similarity and other issues are discussed in Sharafoddini et al. 2017 [11]. Two methods for defining similarity are a simple Euclidean distance and cosine similarity. Park et al[[8] was able to show that an Euclidean distance metric while carefully investigating the optimum number of neighbors for each patient was able to outperform several conventional machine learning algorithms, including logistic regression, and decision tree classifications. Similarly, patient similarity analytics using cosine-similarity metrics have been demonstrated to outperform population-based models and well-known clinical scoring systems [11].

1.3 Data

Publicly available data from the PhysioNet Computing in Cardiology 2019 competition was sourced from ICU patients in two hospital systems involving 40,336 patients, 2932 of which develop sepsis within their ICU stay (7.3% incidence) [9] [10]. Each patient's data includes hourly measurements of 40 variables, comprising demographics, vital signs and laboratory values and a binary variable of sepsis status. Sepsis status is defined by the first clinical suspicion of infection (later confirmed) or the occurrence of organ damage identified by a two-point increase in Sequential Organ Failure Assessment (SOFA) score within 24 hours. Sepsis labels were pushed forward 6 hours in order to facilitate the prediction of sepsis 6 hours before onset. Covariates in the data are able to describe observations of the patient and administrative variables which reflect care decisions made by clinicians. Covariates are described by table **??**.

2 Methods

2.1 Patient training and test data

For the purpose of our research question, patients had to meet an inclusion criterion: having data records for the first 6 hours of their ICU stay (omitting 643 patients). Only the first 6 hours of their ICU stay were used for analysis. Patients who developed sepsis before being in the ICU for at least 12 hours were also excluded (8728 cases removed).

One binary response variable was created for each patient. A patient was labeled as positive for sepsis if they developed it in the 12th hour or beyond of being in the ICU and negative if they never developed sepsis.

The caret package [4] was used for several pre-processing step including creating a train and test set of data before selecting predictors. Data was first stratified based on class and then sampled within each class to preserve the class balance of the original dataset.

Predictor variables were extracted from the first 6 hours of longitudinal data in the ICU. The mean, minimum, maximum and standard deviation was calculated for each vital sign of a patient (rows 1-8 of table **??**). Laboratory variables were less commonly measured in the first 6 hours of a patient's ICU and hence displayed a high degree of missingness. Laboratory values are often more invasive to measure and are thus more rare. Clinicians will elect to measure these variables based on the current status of the patient and therefore the decision to test may carry predictive power. Based on this reasoning, we decided to test the inclusion of count variables—the number of times a given lab was taken in the first 6-hour period—as it would reflect the opinion and decisions taken by clinicians. Demographic variables (Age, Gender, Unit type) were also included as predictors.

Predictor variables were then filtered before model training. Predictors were removed if they were over 20% missing and incomplete patient records were subsequently removed. Predictors were also removed if they exhibited zero variance or correlation values over 0.9.

2.2 Initial Model Fitting

Initial models for sepsis prediction were trained on the complete population of ICU admissions after modification explained in the prior section. Our goal here was to optimize a baseline model for sepsis prediction, as well as giving the oppourtunity to explore initial modeling questions such as whether to include lab counts in the model, explore feature importance, and consider modifications to the procedure.

Two types of predictive models were assessed using 20 rounds of 10-fold cross validation including logistic regression (LR) and random forest (RF). All models treated sepsis as a binary class problem. Performance of final models was evaluated by calculating the area under the receiver operating curve (AUROC), sensitivity, specificity and area under the precision-recall curve (AUPRC). AUPRC is an informative performance measure for skewed dataset, such as the dataset being investigated here.

In order to alleviate potential effects of the skew class distribution, the cross-validation procedure incorporated a stratified sampling step in order to preserve the skewed class balance into each of the folds. This was to ensure each model developed within cross validation was trained and tested data with the same ratio of sepsis to non-sepsis cases.

The addition of downsampling into the model procedure was also evaluated to help mitigate the effect of class imbalance. Down-sampling refers to the practice of sampling within the majority class to decrease its size, allowing it to be equally balanced with the minority class while the minority class remains unchanged. It typically occurs as a preprocessing step before cross validation however this approach can give over-optimistic and less robust estimates of error [4]. Here, downsampling was instead implemented to occur

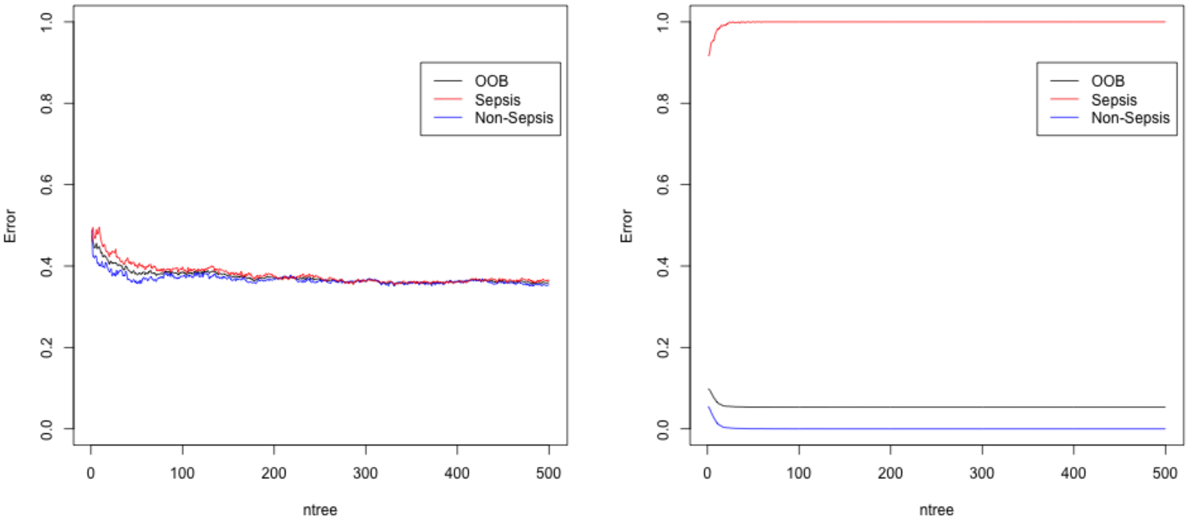

Figure 1: Error Rates of Random Forests trained on lab counts as trees are grown for down sampled (left) and normal (right) cross validation.

within the training resamples of cross validation in order to be treated as a component of the procedure.

The optimal population-level models were then tested on the final hold-out test data.

2.2.1 Logistic Regression

The probability of a patient developing sepsis is modeled using a simple Logistic Regression:

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \theta^T \mathbf{X_i}$$

$$\pi_i = P(S_i = 1 | P_i = p_i)$$

$$\mathbf{P_i} = [1, P_{i1}, .., P_{ip}]^T$$

where $S = 1$ indicates a patient has developed Sepsis, 0 otherwise and P is the vector of predictor variables.

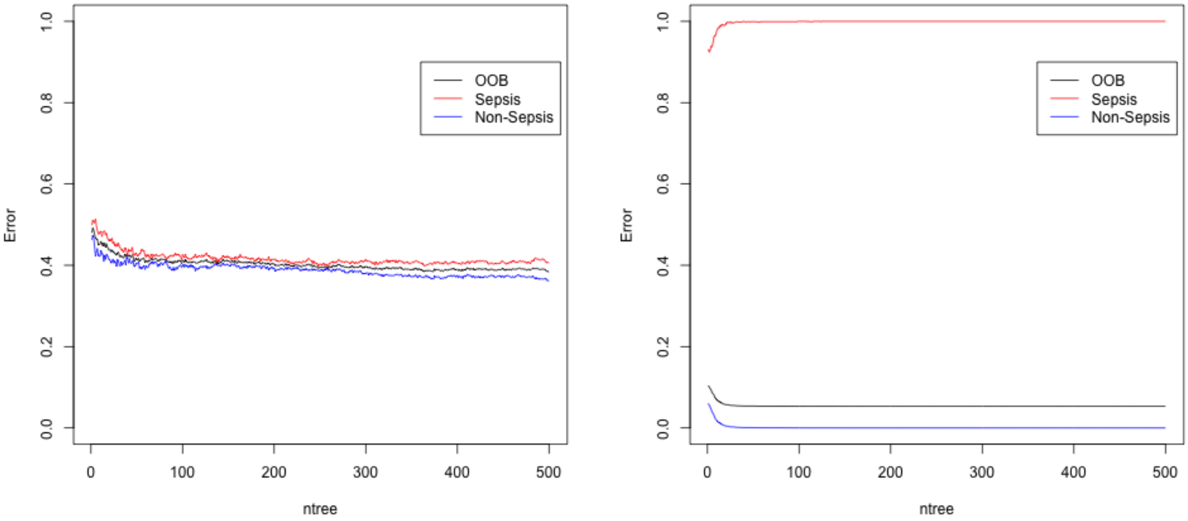

Figure 2: Error Rates of Random Forests trained on only vital signs as trees are grown for down sampled (left) and normal (right) cross validation.

2.2.2 Random Forest

The probability of a patient developing sepsis is modeled using a random forest as implemented in the original publication[?]. Here, the estimated probability is the proportion of trees which voted for the Sepsis class.

Random Forests were tuned using a grid search, optimizing for AUROC. The number of covariates randomly sampled as candidates for each split were selected from (3, 4, 5, 6, 7, 8, 10, 15, 16, 17) and the number of trees to grow selected was 100, 200 or 500. By cross-validation results, 3 was determined to be the optimal number of covariates randomly sampled as candidates for each split. The error rate of each random forest was shown to stabilize before 200 trees were grown in figures 2 and 1. Random forests with less trees are more computationally favourable so 200 trees and 3 candidates for splits was selected for the baseline similarity model.

2.3 Patient Similarity Models

After initial model testing and tuning, new personalized models were trained based on similar patients, using the optimal parameters and predictor variables from initial cross validation.

For a personalized model based on similar patients, the set of a patient's predictor variables is represented as an Euclidean vector, P_i. Two patient similarity metrics (PSM) were selected for this analysis, the Euclidean distance and the cosine similarity, which is the cosine of the angle between the two vectors.

The Euclidean distance is based on the L2 distance of the two vectors and is mathematically

$$PSM_1 = \sqrt{||P_1 \cdot P_2||}$$

Where P_1 and P_2 are the patient vectors. Values range from 0 to infinity where patients have the same characteristics with distance 0 and increases with the value of distance.

Cosine similarity is based on the similarity of orientation of two vectors, taking the cosine of the angle,

$$PSM_2 = \frac{P_1 \cdot P_2}{||P_1||||P_2||}$$

Values range from -1 to 1, where -1 means exact opposite orientation, (i.e. 180 degrees), 1 meaning the exact same (i.e. 0 degrees), and 0 being orthogonal (i.e. 90 degrees).

2.4 Training and Evaluation

Each personalized predictive model was validated with 20 rounds of 10-fold cross validation. For each patient in a test or validation set as the index patient, the following steps were completed,

1. The patient similarity metric is calculated for every patient in the training set relative to the index patient.
2. The PSM values are sorted in ascending order for Euclidean distance and descending order for cosine similarity.

3. The top N most similar patients are utilized for the training data for testing in the validation set.

4. Each model predicts the sepsis risk of the index patient, the probability between 0 and 1 of developing sepsis. If all of the patients in the similarity cohort had developed sepsis or none had, the probability assigned would be 1 or 0, respectively.

Including downsampling in the procedure was also considered and tested for each model type. Within a cross validation training fold, patients would first be down sampled to obtain an equal class balance before the above steps would be taken, effectively diminishing the pool of similar patients. The class balance would remain skewed in the testing fold.

N was varied between 100 and 1900 for the procedure which included downsampling, 1940 being the minimum size of a down-sampled training fold while N was varied between 100 and 17000 in the original procedure.

Patient Similarity models were also assessed using AUPRC and AUROC. Algorithms in the competition were assessed using a unique performance metric which penalized late predictions and false alarms [7]. However, this utility function was not used here because it was designed for algorithms which made hourly predictions of the patient's sepsis 6 hours in the future and therefore had to consider early or late predictions. AUROC was selected because it equally considers specificity and sensitivity, the true negative rate and true positive case, and thus it will penalize false positives.

The optimal patient similarity models were then tested on the final hold-out test data.

3 Results

20,281 ICU admissions in the PhysioNet challenge had complete data from the first 6 hours in the ICU and were used for model training. Table 1 describes the patient characteristics. The incidence rate of sepsis was 5.3%.

3.1 Predictive Performance of Population-level Models

To build baseline predictive models, provide a benchmark of success and answer initial modeling questions, population-level models of Logistic Regression and Random Forest were trained, tuned (if applicable) and assessed using 10-fold cross-validation.

Each LR and RF model was tuned 4 times on a combination of lab-count and downsampled datasets. The performance of each is described in the tables 3 and 2, using AUROC, as well as Sensitivity, Specificity and AUPRC with grid-search results for random forest shown in table 4.

The optimal logistic regression achieved an AUROC of 0.673 (95% confidence interval (CI): [0.664, 0.672]) and an AUPRC value of 0.11. This baseline LR model including laboratory counts and no downsampling in the model procedure. The optimal random forest achieved an AUROC of 0.687 (95% CI: [0.684, 0.691]) and an AUPRC of 0.11 with 500 trees and 3 candidates at each split. However, the baseline RF model was selected to have 200 trees to ease the computational burden of the similarity procedure. This model with an AUROC of 0.683 [0.68, 0.687] and AUPRC of 0.11 had comparable performance to the fill RF model with 500 trees. The baseline RF model also included laboratory counts and downsampling in the model procedure.

Despite it being the best performing LR model, the logistic regression demonstrated a low sensitivity compared to the corresponding logistic regression with downsampling. Each of the models without downsampling exhibited a similar trend where sensitivity is low (¡ 0.01) and specificity is high (¿0.99). A low or 0 sensitivity indicates few or 0 patients who develop sepsis as being correctly predicted as developing sepsis, instead receiving a negative prediction. Combined with a high specificity where the majority of patients who are healthy are correctly predicted as healthy. This pattern is a result of class imbalance demonstrated in the training set: these models predicted the majority class (non-sepsis) for almost the full testing set, and since the testing sets had a majority of non-sepsis cases, the models still obtained favourable AUROC values. In the case of the random forest models, this pattern is additionally reflected by smaller AUPRC values for models without downsampling. The addition of downsampling in the model procedure improves the ability of each predictive model to correctly identify future sepsis patients.

The addition of lab counts into the population-level LR and RF models also improves predictive performance. This is reflected by a 4-5 percentage point increase in AUROC value and 1-2 percentage point increase in AUPRC as can be seen in tables 3 and 2.

3.1.1 Variable Importance

The variable importance for the top 20 features for the best performing LR and RF population models are presented in figure . We can observe that in the variable importance for logistic regression, only 4 features have a higher importance estimate than 50%, while random forest has 19. This may reflect a number of features having a non-linear relationship with the probability of developing sepsis that can be accommodated in a random forest, yet it cannot be modelled in a more restricted logistic regression model. It's also notable that few features are moderately important (¿50%) across both models, a rare example being mean heart rate and count of labs for Fraction of inspired oxygen (%) (FiO2_count).

3.2 Personalized Predictive Models based on Similar patients with Downsample

Figure ??and ??illustrate the AUROC of logistic regression and random forest, respectively, as a function of the number of similar patients used as training data. We observe for 3 of 4 plots that AUROC begins small and increases with the number of similar patients, peaking around 1500 patients with AUROC similar to that of the population models. This trend shows the trade-off of similarity models, that when too few patients are included, the model will begin to suffer from small-sample size effects. This is a hazard combining downsampling and patient similarity, procedures which both decrease sample size. Figure ?? (left) shows a more problematic effect, where AUROC has decreased greatly from the population level models and does not recover. This could be due to small sampling size where subsampling lead to a group of patients with less predictive power, or insufficient variability within the down samples. Nevertheless, it indicates a pitfall of this procedure.

3.3 Personalized Logistic Regression based on Similar patients

ncomplete data

Figure ?? illustrates the AUROC of logistic regression as a function of the number of similar patients utilized for model training. We observe that for Euclidean distance, AUROC begins small and increases quickly, able to match the performance of the population level model.

3.4 Personalized Random Forest based on Similar patients

Incomplete data.

Incomplete data – parts to include later What were the peak performances and did they outperform the model that used all the available training data.

Of the Logistic regression models, logistic regression with the full dataset including labcounts without downsampling performed the best with an AUROC of 0.6729.

Of the Random Forest models, analysis with the full dataset including lab counts with downsampling in the model procedure performed the best with an AUROC of 0.6871. The random forest model's best tune was with 500 trees and 3 mtry or variables selected.

4 Discussion and Conclusion

In this paper, we present an approach for early sepsis prediction in the ICU based on patient similarity models. Patient similarity demonstrates the intuitive and common-sense practice of basing clinical prognosis on similar cases by training machine learning models on a subset of similar patients. First by building population-level models, we were able to address initial modelling problems such as variable inclusion and class imbalance. Then with the addition of patient similarity, we were able to refine broader models to focus on similar patients which gave promising(?) results. Statistical learning methods driven by patient similarity yielded improved (?) predictive performance compared to population-level models. Our testing of similarity models also validated previous results that showed predictive performance would suffer if the number of similar patients selected was too small.

None the less, there are few areas in this paper that speak to the limitations we encountered. First, A notable limitation is the processing power which is required of a personalized predictive model based on similar patients. Although, the patient similarity metrics used in this paper were computationally efficient, individual models would need to be built for each index patient, several times when conducting cross validation and tuning the model. Despite these functions being easily parallelizable, models would still take a number of days to tune, much longer than a population-based model. These models would have a larger computational burden if the PSM calculations were made more complex. Second, highly missing variables and patients with missing data were excluded from the analysis in this paper. The latter potentially leading to selection bias. Third, a notable difference in the methodology here and in papers with similar methodology [?] is the rescaling of predictor variables between -1 and 1 before calculating PSM values. This is a step to ensure predictor variables have an equal contribution to the PSM calculation and a methodology without it will have a different outcome. To explore the outcome of this effect, a small study of it is included here to test how different the similar patients would be, given rescaled predictor variables and unscaled predictor variables.

Fig ?? illustrates a simulation of patient selection on scaled predictors and unscaled predictors using both cosine similarity (left) and Euclidean distance (right). Both figures represent the percentage of shared patients by both a scaled and unscaled procedure as a function of the total number of similar patients. It can be observed that both plots

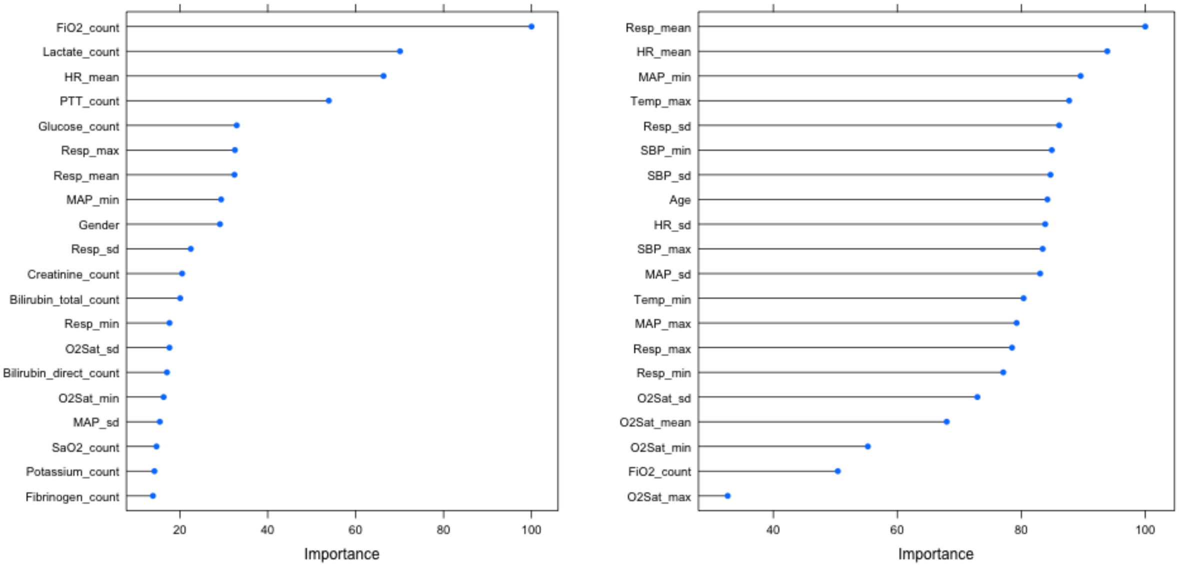

Figure 3: Variable Importance for Classification with best model of Logistic Regression (left) and Random Forest (right).

exhibit a similar trend, that as the neighbourhood of similar patients grow, more patients are shared between procedures. This is most likely due to the depletion of candidate patients. If two samples are taken from the same population and are greater than half the population size, by necessity, there will be overlap. Both similarity metrics will have less than 10% shared by procedure for very small numbers of similar patients. Neither patient selection by cosine similarity or Euclidean distance appears to be invariant to the effect of rescaling predictor variables. Therefore, rescaling will have a notable effect on predictions and hence predictive performance of the similarity models, negatively correlated in size with the number of similar patients.

This paper also leads to a few areas that can spark future work. The aim of the Computing in Cardiology Challenge 2019 [10] was to develop algorithms which would involve a new prediction of future sepsis status every hour in the ICU. The algorithm presented in this paper involves a static one-time prediction of sepsis status beyond 12 hours, given the initial 6 hours of ICU data. However, a similar algorithm could be developed incorporating a survival analysis to enable hourly prediction.

2

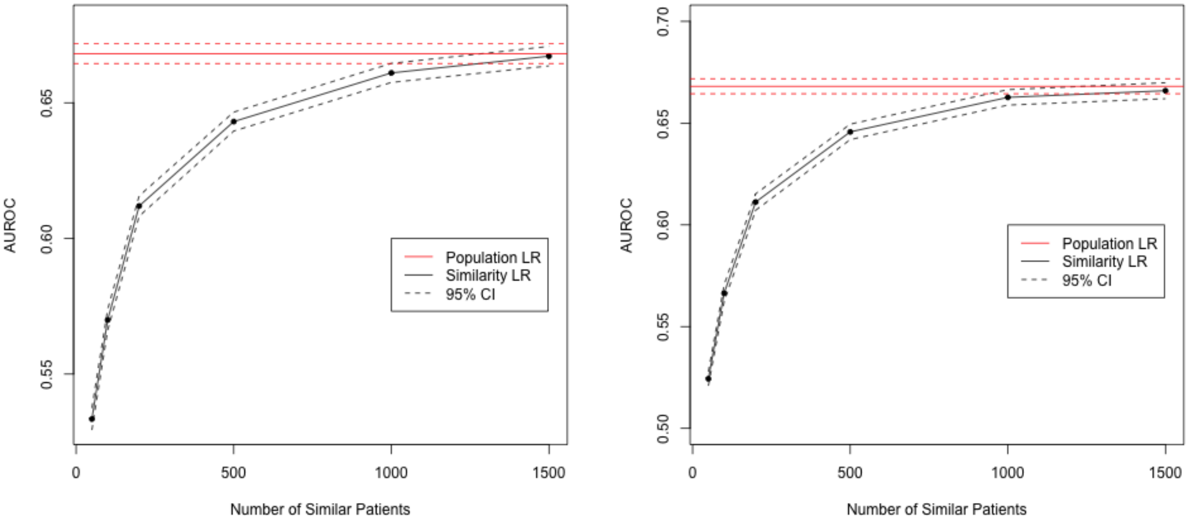

Figure 4: AUROC as a function of Number of Similar patients for Logistic Regression with downsampling. Cosine Similarity is used on the left, Euclidean distance on the right.

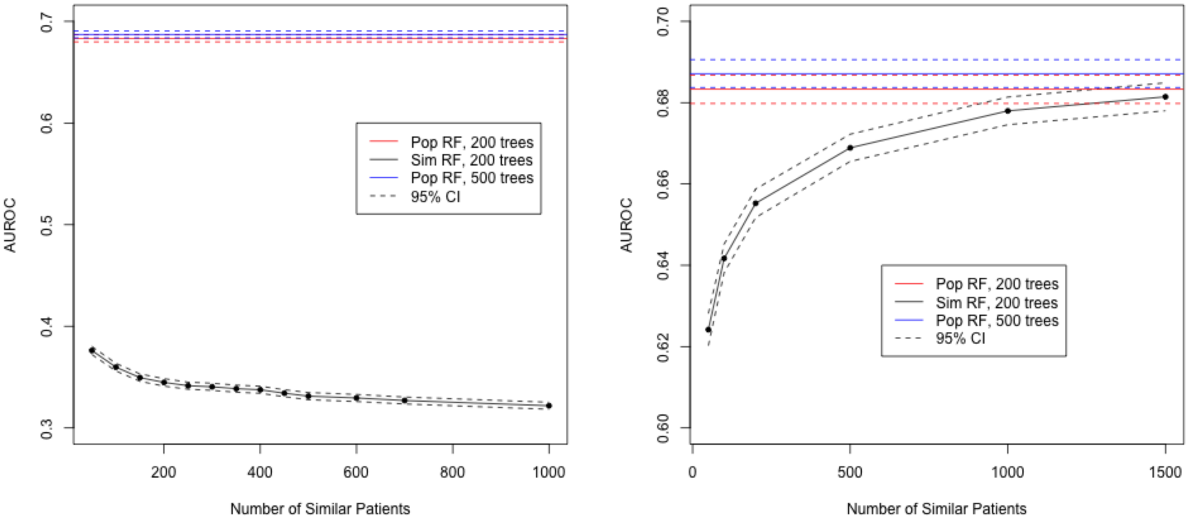

Figure 5: AUROC as a function of Number of Similar patients for Random Forest with downsampling. Cosine Similarity is used on the left, Euclidean distance on the right.

Vital Signs	
HR	Heart rate (beats per minute)
O2Sat	Pulse oximetry (%)
Temp	Temperature (Deg C)
SBP	Systolic BP (mm Hg)
MAP	Mean arterial pressure (mm Hg)
DBP	Diastolic BP (mm Hg)
Resp	Respiration rate (breaths per minute)
EtCO2	End tidal carbon dioxide (mm Hg)
Laboratory values	
BaseExcess	Measure of excess bicarbonate (mmol/L)
HCO3	Bicarbonate (mmol/L)
FiO2	Fraction of inspired oxygen (%)
pH	N/A
PaCO2	Partial pressure of carbon dioxide from arterial blood (mm Hg)
SaO2	Oxygen saturation from arterial blood (%)
AST	Aspartate transaminase (IU/L)
BUN	Blood urea nitrogen (mg/dL)
Alkalinephos	Alkaline phosphatase (IU/L)
Calcium	(mg/dL)
Chloride	(mmol/L)
Creatinine	(mg/dL)
Bilirubin_direct	Bilirubin direct (mg/dL)
Glucose	Serum glucose (mg/dL)
Lactate	Lactic acid (mg/dL)
Magnesium	(mmol/dL)
Phosphate	(mg/dL)
Potassium	(mmol/L)
Bilirubin_total	Total bilirubin (mg/dL)
TroponinI	Troponin I (ng/mL)
Hct	Hematocrit (%)
Hgb	Hemoglobin (g/dL)
PTT	partial thromboplastin time (seconds)
WBC	Leukocyte count (count$*10^3/\mu$ L)
Fibrinogen	(mg/dL)
Platelets	(count$*10^3\mu$L)
Demographics	
Age	Years (100 for patients 90 or above)
Gender	Female (0) or Male (1)
Unit1	Administrative identifier for ICU unit (MICU)
Unit2	Administrative identifier for ICU unit (SICU)
HospAdmTime	Hours between hospital admit and ICU admit
ICULOS	ICU length-of-stay (hours since ICU admit)
SepsisLabel	For sepsis patients, SepsisLabel is 1 if \$t +6> t_sepsis\$, 0 otherwise.

Number of unique, complete, ICU patients	**20,281**
Age	61.6 [16.7]
Gender (% Male)	54.6
Unit Type	
Medical Intensive Care Unit (%)	33.7
Surgical Intensive Care Unit (%)	28.0
Sepsis Incidence (%)	**5.315**

Table 1: Patient data characteristics in the Training Set. Age is shown in mean [standard deviation].

Model Type	AUROC	Sens	Spec	AUPRC
Lab Counts, Down	0.668 [0.664, 0.672]	0.577 [0.57, 0.584]	0.677 [0.675, 0.679]	0.11
Lab Counts	0.673 [0.669, 0.677]	0.004 [0.003, 0.005]	1 [1, 1]	0.11
Only Vitals, Down	0.625 [0.621, 0.629]	0.567 [0.56, 0.573]	0.614 [0.613, 0.616]	0.09
Only Vitals	0.629 [0.625, 0.633]	0 [0, 0]	1 [1, 1]	0.09

Table 2: Performance of Population-level Logistic Regression. Shown in mean [95% CI].

Model Type	AUROC	Sens	Spec	AUPRC
Lab Counts, Down	0.687 [0.684, 0.691]	0.639 [0.633, 0.645]	0.64 [0.638, 0.642]	0.11
Lab Counts	0.678 [0.674, 0.681]	0 [0, 0]	1 [1, 1]	0.1
Only Vitals, Down	0.639 [0.635, 0.642]	0.604 [0.597, 0.611]	0.599 [0.597, 0.601]	0.09
Only Vitals	0.63 [0.626, 0.633]	0 [0, 0]	1 [1, 1]	0.09

Table 3: Performance of Population-level Random Forest. Shown in mean [95% CI].

Model Type	Number of Trees	Number of Candidates at each Split
Lab Counts, Down	500	3
Lab Counts	500	3
Only Vitals, Down	500	4
Only Vitals	500	3

Table 4: Parameters of Population-level Random Forest which achieved optimal AUROC.

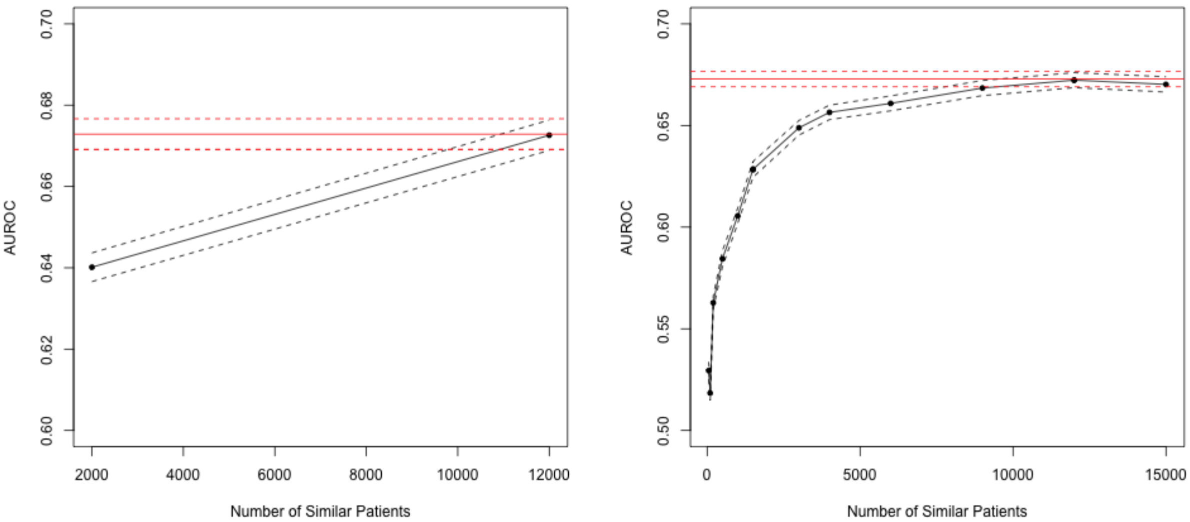

Figure 6: AUROC as a function of Number of Similar patients for Logistic Regression. Cosine Similarity is used on the left, Euclidean distance on the right.

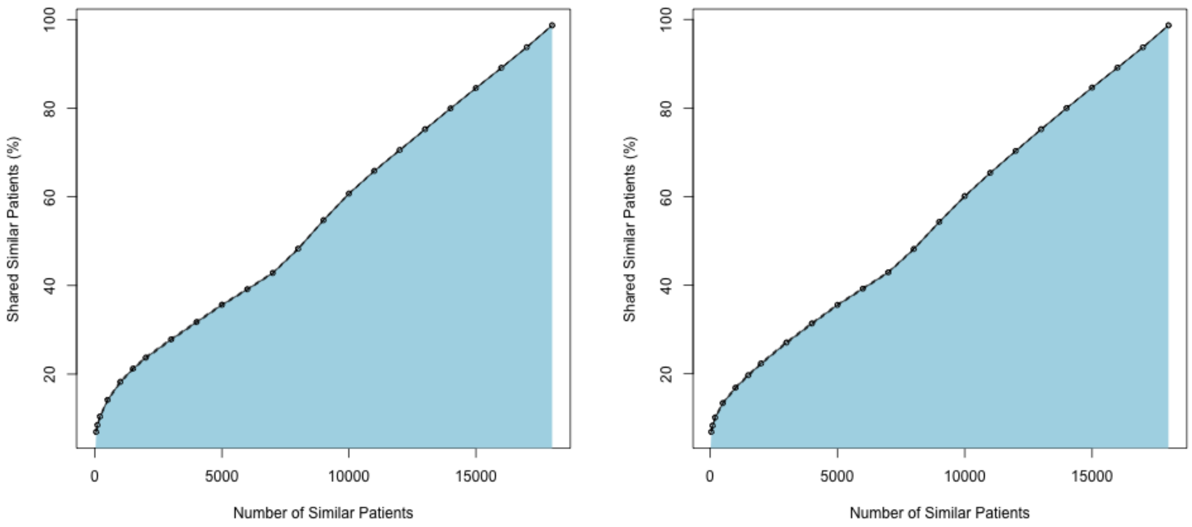

Figure 7: Number of Shared Patients Across Scaled Predictors and Unscaled predictors as a function of Number of Similar patients. Cosine Similarity is used on the left, Euclidean distance on the right.

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[3] Liudmila Husak, Annette Marcuzzi, Jeremy Herring, Eugene Wen, Ling Yin, Dragos Daniel Capan, and Geta Cernat. National analysis of sepsis hospitalizations and factors contributing to sepsis in-hospital mortality in canada. *Age*, 37:39–7, 2010.

[4] Max Kuhn. The caret package, 2009.

[5] Joon Lee, David M. Maslove, and Joel A. Dubin. Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PLOS ONE*, 10(5):1–13, 05 2015.

[6] Andrew Lever and Iain Mackenzie. Sepsis: definition, epidemiology, and diagnosis. *Bmj*, 335(7625):879–883, 2007.

[7] H. Bryant Nguyen, Emanuel P. Rivers, Fredrick M. Abrahamian, Gregory J. Moran, Edward Abraham, Stephen Trzeciak, David T. Huang, Tiffany Osborn, Dennis Stevens, and David A. Talan. Severe sepsis and septic shock: Review of the literature and emergency department management guidelines. *Annals of Emergency Medicine*, 48(1):54.e1, 2006.

[8] Yoon-Joo Park, Byung-Chun Kim, and Se-Hak Chun. New knowledge extraction technique using probability for case-based reasoning: application to medical diagnosis. *Expert systems*, 23(1):2–20, 2006.

[9] Josef Chris Jeter Russell Shashikumar Supreeth Moody Benjamin Westover M. Brandon Sharma Ashish Nemati Shamim Reyna, Matthew and Gari Clifford. Early prediction of sepsis from clinical data – the physionet computing in cardiology challenge 2019. *PhysioNet*, (1.0.0), 2019.

[10] Matthew A. Reyna, Christopher S. Josef, Russell Jeter, Supreeth P. Shashikumar, M. Brandon Westover, Shamim Nemati, Gari D. Clifford, and Ashish Sharma. Early prediction of sepsis from clinical data: The physionet/computing in cardiology challenge 2019. *Critical Care Medicine*, 48(2), 2020.

[11] Anis Sharafoddini, Joel A Dubin, and Joon Lee. Patient similarity in prediction models based on health data: a scoping review. *JMIR medical informatics*, 5(1):e7, 2017.

[12] 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). *Jama*, 315(8):801–810, 2016.

[13] Hadley Wickham and Garrett Grolemund. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, Inc., 1st edition, 2017.