Early prediction of patient mortality based on routine laboratory tests and predictive models in critically ill patients

Sven Van Poucke
Ziekenhuis Oost Limburg
96 PUBLICATIONS **254** CITATIONS

SEE PROFILE

Ana Kovacevic
3 PUBLICATIONS **9** CITATIONS

SEE PROFILE

Milan Vukicevic
University of Belgrade
38 PUBLICATIONS **196** CITATIONS

SEE PROFILE

Early Prediction of Patient Mortality Based on Routine Laboratory Tests and Predictive Models in Critically Ill Patients

Sven Van Poucke, Ana Kovacevic and
Milan Vukicevic

Additional information is available at the end of the chapter

Abstract

We propose a method for quantitative analysis of predictive power of laboratory tests and early detection of mortality risk by usage of predictive models and feature selection techniques. Our method allows automatic feature selection, model selection, and evaluation of predictive models. Experimental evaluation was conducted on patients with renal failure admitted to ICUs (medical intensive care, surgical intensive care, cardiac, and cardiac surgery recovery units) at Boston's Beth Israel Deaconess Medical Center. Data are extracted from Multi parameter Intelligent Monitoring in Intensive Care III (MIMIC-III) database. We built and evaluated different single (e.g. Logistic regression) and ensemble (e.g. Random Forest) learning methods. Results revealed high predictive accuracy (area under the precision-recall curve (AUPRC) values >86%) from day four, with acceptable results on the second (>81%) and third day (>85%). Random forests seem to provide the best predictive accuracy. Feature selection techniques Gini and ReliefF scored best in most cases. Lactate, white blood cells, sodium, anion gap, chloride, bicarbonate, creatinine, urea nitrogen, potassium, glucose, INR, hemoglobin, phosphate, total bilirubin, and base excess were most predictive for hospital mortality. Ensemble learning methods are able to predict hospital mortality with high accuracy, based on laboratory tests and provide ranking in predictive priority.

Keywords: mortality risk prediction, renal failure, metabolic panel, feature selection, ensemble methods

1. Introduction

Precision medicine is based on comprehensive models with the potential to elucidate the complexity of health and diseases, including the features of emergence, nonlinearity, self-organization, and adaptation [1].

Laboratory testing is more common among patients admitted to ICU [2, 3]. Blood sample frequencies vary, but routinely tests are ordered by fixed schedule and in clusters as part of the hypothetico-deductive diagnostic exploration. Quantitative predictive analysis of daily sampling might provide new insights into the choice (feature selection) and importance (feature weighting) of each laboratory test [4]. In this chapter, we propose a system for mortality risk prediction of patients with renal failure, based on predictive methods. Renal failure patients were selected based on the Elixhauser Comorbidity Index [5]. For chronic disease, the use of Elixhauser is sensitive for the systemic underrepresentation of chronic conditions [6, 7].

This study quantitatively assessed the predictive power of laboratory tests for hospital mortality in patients admitted to ICU. Based on previous findings, we compared the predictive performance of different single (Decision Tree, Naive Bayes, Logistic and Regression) and ensemble (Random Forest, Boosting, and Bagging) learning methods. Moreover, the predictive power and importance of predictors (laboratory tests) were quantitatively assessed by use of feature weighting and selection techniques: Correlation, Gini Selection, Information Gain and ReliefF [8]. For predictive modeling, feature selection, and visual analytics of the results, we used RapidMiner and R platforms as mentioned in [9–11].

2. Materials and methods

2.1. Data source and study subjects

The MIMIC-III (version 1.0) clinical database consists of 58,976 ICU admissions for 46,520 distinct patients, admitted to Beth Israel Deaconess Medical Center (Boston, MA) from 2001 to 2012 [12, 13]. The establishment of the database was approved by the Institutional Review Boards of the Massachusetts Institute of Technology (Cambridge, MA) and Beth Israel Deaconess Medical Center (Boston, MA). Accessing the database was approved for authors S.V.P and Z.Z. (certification number: 1712927 and 1132877). Informed consent was waived due to observational nature of the study.

The MIMIC-III clinical database includes data related to patient demographics, hospital admissions and discharge dates, room tracking, death dates (in or out of the hospital), ICD-9 codes, health care providers, and types. All dates were surrogate dates but time intervals were preserved. In addition, physiological data, medications consumption, laboratory investigations, fluid balance calculations and notes, and reports were included in the basic dataset.

2.2. Data preparation

RapidMiner was used because it enabled handling unstructured data without the need for coding [9, 14].

The dataset in this study was generated by joining data from the following MIMIC-III tables: admission, patients, ICU stays, diagnoses_icd, and lab events. Patients were assigned to sub-populations including hypertension, paralysis, chronic pulmonary disease, diabetes, renal failure, acquired immunodeficiency syndrome (AIDS), coagulopathy, obesity, and weight loss, and so on, based on the Elixhauser comorbidity score [6, 7]. Renal failure is defined in the Elixhauser comorbidity score, when ICD-9 code is in (70.32, 70.33, 70.54, 456, 456.1, 456.2, 456.21, 571, 571.2, 571.3, ≥ 571.4 ≤ 571.49, 571.5, 571.6, 571.8, 571.9, 572.3, 572.8, and 42.7).

All time stamped measurements in MIMIC-III were zeroed in reference to the moment of hospital admission.

2.3. Predictive algorithms

The process compared different learning and ensemble methods (Decision Stump, Decision Tree, Naive Bayes, Logistic Regression (LR), Random Forest, Support Vector Machine, AdaBoost, Bagging, and Stacking) in association with feature weighting and selection, quantitatively assessed in terms of Correlation, Gini Selection, and Information Gain and ReliefF as previously described [8].

2.3.1. Single learning methods

Decision trees (DT) are predictive algorithms based on “greedy,” top-down recursively partitioning of data. DT algorithms perform an exhaustive search over all possible splits in every recursive step. The attribute (predictor) demonstrating the best split by an evaluation measure selected for branching the tree. Regularly used are information theoretic measures (e.g. Information Gain, Gain Ratio, Gini, etc.) or statistical tests quantifying the significance of the association between predictors and class. The procedure is recursively iterated until a stop criterion is met [15, 16]. In this research, we used the J48 algorithm, which is the Java implementation of the C4.5 algorithm [17].

Logistic regression (LR) is a linear classifier modeling the probability of a dependent binary variable y given a vector of independent variables X. For the estimation of the probability, the example belongs to the positive class, a logit model is used:

$$log\left(\frac{p}{1-p}\right) = \Theta_0 + \Theta_1 x_1 + \cdots + \Theta_n x_n \tag{1}$$

where p presents probability that y = 1, θj, j = 1,…,n present the weights of the corresponding dependent variable, while p/(1-p) is called odds ratio, parameters θj, j = 1,…,n of the model can be interpreted as changes in log odds or the results can be interpreted in terms of probabilities [18–20].

2.3.2. Ensemble learning methods

Ensemble (meta-learning) methods combine multiple models aiming to provide more accurate or more stable predictions. These models can be aggregated from the same model built on different sub-samples of data, from different models built on the same sample or a combination of the previous two techniques. Ensemble methods are often used to improve the individual performance of algorithms that constitute ensembles by exploiting the diversity among the models produced [21]. The ensemble methods implemented in this chapter are: Random Forest [22], Boosting [23], and Bootstrap Aggregating (Bagging) [24]. In our experiments, Boosting and Bagging used J4.8 and Logistic regression as base learners.

Random Forest (RF) is an ensemble classifier that evaluates multiple DT and aggregates their results, by majority voting, in order to classify an example [22]. There is a two-level randomization in building these models. First, each tree is trained on a bootstrap sample of the training data and second, in each recursive iteration of building a DT (splitting data based on information potential of features); a subset of features for evaluation is randomly selected. In this research, we grew and evaluated Random Forest (RF) with 10 trees.

Boosting is an ensemble meta-algorithm developed in order to improve supervised learning performance of weak learners (models whose predictive performance is only slightly better than random guessing). In this study, the adaptive boosting (AdaBoost) algorithm was used [23].

Bagging algorithm builds a series of models (e.g. CHAID Decision Trees) on different data subsamples (with replacement) [24]. For new examples, each model is applied, and predictions are aggregated (e.g. majority voting for classification or average prediction for regression).

2.4. Feature weighting and selection

Several filter feature selection schemes were evaluated. Filter selection (FS) methods rely on the evaluation of the information potential of each input feature in relation to the label (hospital mortality). A threshold search and selection of those features, providing most predictive power, was calculated for each predictive model. The first is based on Pearson correlation returning the absolute or squared value of the correlation as attribute weight. Furthermore, we applied Information Gain Ratio and Gini Index, two weighting schemes that are based on information theoretic measures, frequently used with decision trees for evaluation of potential splits [17]. The T-test calculated, for each attribute, a p-value for two-sided, two-sample T-test. Finally, the ReliefF evaluated the impact of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class [25].

3. Experimental evaluation

3.1. Exploratory analyses

The MIMIC-III database consists of 58,976 hospital admissions from 46,520 patients. All patients are characterized by at least one ICU admission.

Guided by the CRoss-Industry Standard Process for Data Mining (CRISP-DM), the ETL process was initiated by retrieving data from the MIMIC-III tables of interest (d_labitems, admissions, patients, and diagnosis_icd) [26]. Next, patients were selected for renal failure by the Elixhauser score, leading to 1477 (3.15%) patients satisfying our inclusion criteria and 20,068 patient days (examples) in total. In a consecutive step, admissions were joined based on hospital admission id (hadm_id) with all laboratory tests (from the 755 item ids in d_labitems), aggregated on a daily level. Mean, standard deviation, and the number of tests per day (*len*) were defined as aggregation functions. As an output feature (*label*), this study focused on hospital mortality (hospital_expire_flag). From all renal failure patients in this study, 399 (27.0%) did not survive during hospital admission. Next, data were split per day in order to examine feature selection and weight changes over time. Therefore, we arbitrarily limited our computations for admission duration of 7 days, where for each day the number of patients was >1000. After that period, the number of patients admitted to ICU declined.

Patients who survived hospital stay were significantly older (69.3 ± 12.4 years vs. 65.9 ± 14.1 years; $p < 0.05$), suffered more frequently from deficiency anemia (15.5 vs. 9.8% $p = 0.01$) and depression (8.3 vs. 3.8% $p = 0.00$). The survivors suffered less frequently from congestive heart failure (40.2 vs. 46.9% $p = 0.02$), valvular disease (9.8 vs. 14.3% $p = 0.01$), lymphoma (1.6 vs. 3.8% $p = 0.01$), and metastatic cancer (1.7 vs. 4.5% $p = 0.00$). **Table 1** displays the basic characteristics of the baseline dataset. Binary variables are reported as prevalence percentages or count, and continuous variables are reported as data mean ± standard deviation.

In **Figure 1**, distributions of numbers of laboratory tests by admission days are described by the box plots for each day demonstrating a decline in the number of different laboratory tests requested by admission days from day 1 to day 4. For the following days, the number of requested laboratory tests was stable.

3.2. Automatic model building, feature selection and evaluation

A more detailed technical description of the use of RapidMiner for scalable predictive analytics of medical data, as well as templates of generic processes, can be found in [8] and its supplementary materials.

Initially, all features are weighted by five feature weighting and selection methods (Information Gain ratio, Gini, Correlation, ReliefF, and T-test), for each day. In order to find the adequate number of features that will be used by each predictive model for each day (and to identify optimal feature selection methods for our data), we conducted the following procedure. First, we sorted the features by their weights in descending order (for each feature weighting method). Then we trained each of five predictive models (Decision tree, Logistic regression, Random Forest, Bagging, and Boosting) on subsets of features with highest weights, starting from 10 features up to 100 with the step of 10 (9 different feature sets) [27, 28]. Even though a number of experiments were conducted (315 experiments: 7 algorithms X 5 feature selection schemes X 9 thresholds), this method as previously described [8] allowed ease of implementation of the experimental setup within only one RapidMiner process execution and with complete reproducibility of the results.

Characteristics	ICU patients with renal failure (n = 1477)	Survival during hospital admission (n = 1078) (73.0%)	Death during hospital admission (n = 399) (27.0%)	p
Age (years)	66.8 ± 13.8	69.3 ± 12.4	65.9 ± 14.1	*<0.05*
Sex (male, %)	60.1	59.4	61.9	
Congestive heart failure	620 (42.0)	433 (40.2)	187 (46.9)	*0.02*
Cardiac arrhythmias	438 (29.7)	307 (28.5)	131 (32.8)	0.10
Valvular disease	163 (11.0)	106 (9.8)	57 (14.3)	*0.01*
Pulmonary circulation	95 (6.4)	68 (6.3)	27 (6.8)	0.75
Peripheral vascular disease	274 (18.6)	199 (18.5)	75 (18.8)	0.88
Hypertension	13 (0.9)	8 (0.7)	5 (1.3)	0.35
Paralysis	19 (1.3)	13 (1.2)	6 (1.5)	0.65
Other neurological	70 (4.7)	46 (4.3)	24 (6.0)	0.16
Chronic pulmonary disease	253 (17.1)	188 (17.4)	65 (16.3)	0.60
Diabetes uncomplicated	322 (21.8)	225 (20.9)	97 (24.3)	0.15
Diabetes complicated	435 (28.5)	326 (30.2)	109 (27.3)	0.27
Hypothyroidism	155 (10.5)	120 (11.1)	35 (8.8)	0.19
Renal failure	1477 (100)	1078 (100.0)	399 (100.0)	
Liver disease	76 (5.1)	51 (4.7)	25 (6.3)	0.24
Peptic ulcer	12 (0.8)	10 (0.9)	2 (0.5)	0.42
Aids	19 (1.3)	14 (1.3)	5 (1.3)	0.94
Lymphoma	32 (2.2)	17 (1.6)	15 (3.8)	*0.01*
Metastatic cancer	36 (2.4)	18 (1.7)	18 (4.5)	*0.00*
Solid tumor	68 (4.6)	46 (4.3)	22 (5.5)	0.31
Rheumatoid arthritis	44 (3.0)	31 (2.9)	13 (3.3)	0.70
Coagulopathy	149 (10.1)	106 (9.8)	43 (10.8)	0.59
Obesity	34 (2.3)	28 (2.6)	6 (1.5)	0.21
Weight loss	60 (4.1)	37 (3.4)	23 (5.8)	*0.04*
Fluid electrolyte	572 (38.7)	414 (38.4)	158 (39.6)	0.68
Blood loss anemia	0 (0.0)	0	0	0
Deficiency anemias	206 (13.9)	167 (15.5)	39 (9.8)	*0.01*
Alcohol abuse	41 (2.8)	29 (2.7)	12 (3.0)	0.74
Drug abuse	25 (1.7)	19 (1.8)	39 (1.5)	0.73
Psychoses	42 (2.8)	32 (3.0)	10 (2.5)	0.63
Depression	105 (7.1)	90 (8.3)	15 (3.8)	*0.00*

Table 1. Patient characteristics of the baseline dataset.

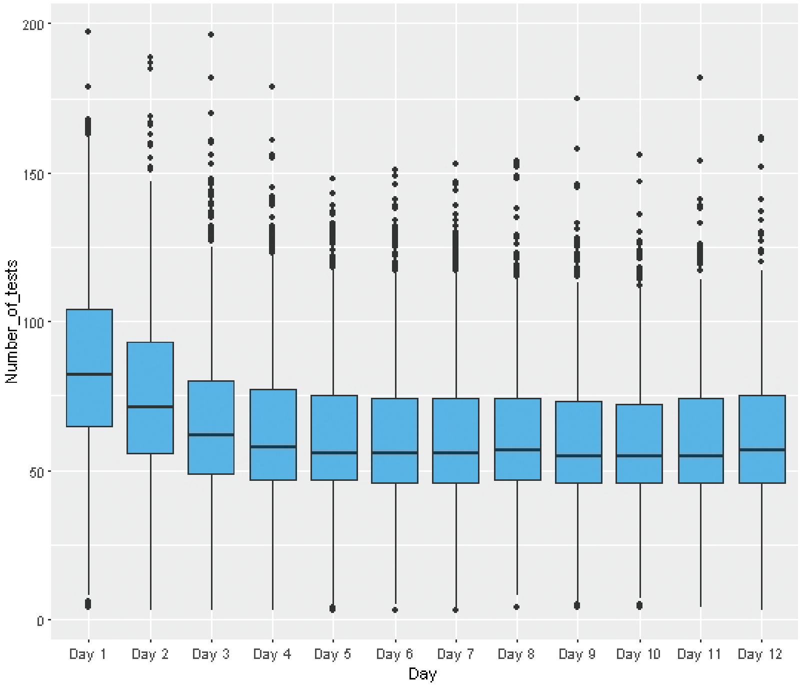

Figure 1. Distribution of the number of laboratory tests per patient by days.

Evaluation of all predictive models was performed by AUPRC (area under the precision-recall curve) for model comparison, because of the unbalanced nature of data [27, 28]. Namely, frequently used area under receiver operating curve (AUROC) is calculated based on true positive rate and false positive rate. True positive rate may be high even if recall is low (situation when predictor rarely predicts positive class), and thus it is often misleading in case of imbalanced data.

Because of relatively small number of samples (between 1000 and 1500 for each day), the predictive performance of models on unseen data was estimated by a fivefold cross-validation set created by stratified sampling, preserving the initial distribution of positive and negative classes of the target attribute. This validation on relatively small samples avoids the risk of misleading interpretation of the results caused by biased selection of a test set based on one sample.

3.3. Performance and feature selection

First, we present a comparison between feature selection methods, based on maximal predictive performance (in terms of AUPRC) overall algorithms. Next, we restrict further analyses on experiments with the overall best feature selection technique. Values in **Table 2** illustrate the maximal predictive performance for each day and for each feature selection technique. Maximum values by days (rows) are shown in bold. It can be seen that Gini and ReliefF achieved maximum values on all days, except for the first day of ICU admission.

Day/FS method	Info Gain Ratio	Gini	Correlation	ReliefF	T-test
1	0.44	0.47	**0.48**	0.45	0.33
2	0.79	**0.84**	0.82	**0.84**	0.77
3	0.80	**0.85**	**0.85**	**0.85**	0.78
4	0.83	**0.86**	**0.86**	**0.86**	0.78
5	0.84	**0.86**	**0.86**	**0.86**	0.80
6	0.83	**0.86**	0.85	**0.86**	0.77
7	0.84	**0.86**	**0.86**	**0.86**	0.77

Bold values represent the best results per rows. Multiple bold Values per row means that there was more than one equally good results.

Table 2. Maximum area under the precision-recall curve (AUPRC) performance of algorithms per days (rows) and feature selection measures (columns).

On the first day, the maximal performance is achieved with correlation (0.48). A more detailed inspection of the model performance and the number of features selected for each admission day demonstrated that Random Forest, result in the best predictive performance overall days (except the first day). Logistic regression often achieved a good performance. J4.8. achieved the worst AUPRC performance over all days, but in synergy with the AdaBoost ensemble scheme, it provided a competitive performance with Random Forest and Logistic Regression.

Further, **Table 2** illustrates that predictive performance is increasing over days and stabilizes from day 4 to 7 on AUPRC = 0.86. AUPRC values for days 2 and 3 are also high (0.83 and 0.85, respectively) and illustrate that risk for hospital mortality can be predicted, with high confidence, starting from the second day of admission.

Values of area under precision-recall curve illustrate the general performance of predictive models but do not explain anything related to the selection of the actual thresholds that should be selected for predictions. Therefore, we analyzed possible thresholds by inspecting precision-recall (PR) trade-off. High recall means that most of the positive examples (in this case, hospital death) are predicted correctly. High precision means that there is a low number of false alarming (mortality is predicted, but the patient survived). **Figure 2** shows precision/recall (PR) curves for first 4 days that were generated from the predictions of the best performing models (**Table 3**), built on features from the Gini selection.

On the first day, all models resulted in poor results which are found on the upper left PR curve in **Figure 2**.

Highest recall can be achieved with 0.3 precision (70% of false alarms), so this model is not useful, regardless of the threshold selection. On days 2–4, maximal recall can be achieved with around 20% of false alarms. Considering the cost of false negative predictions (low recall), we argue that optimal threshold values for day 2–4 models should be between 0.8 and 1 of recall.

Further, rankings of features provided by Gini feature selection methods illustrated in **Table 3** demonstrate that different algorithms achieved best predictive accuracies with different numbers of selected features. The number of features selected varied over the days.

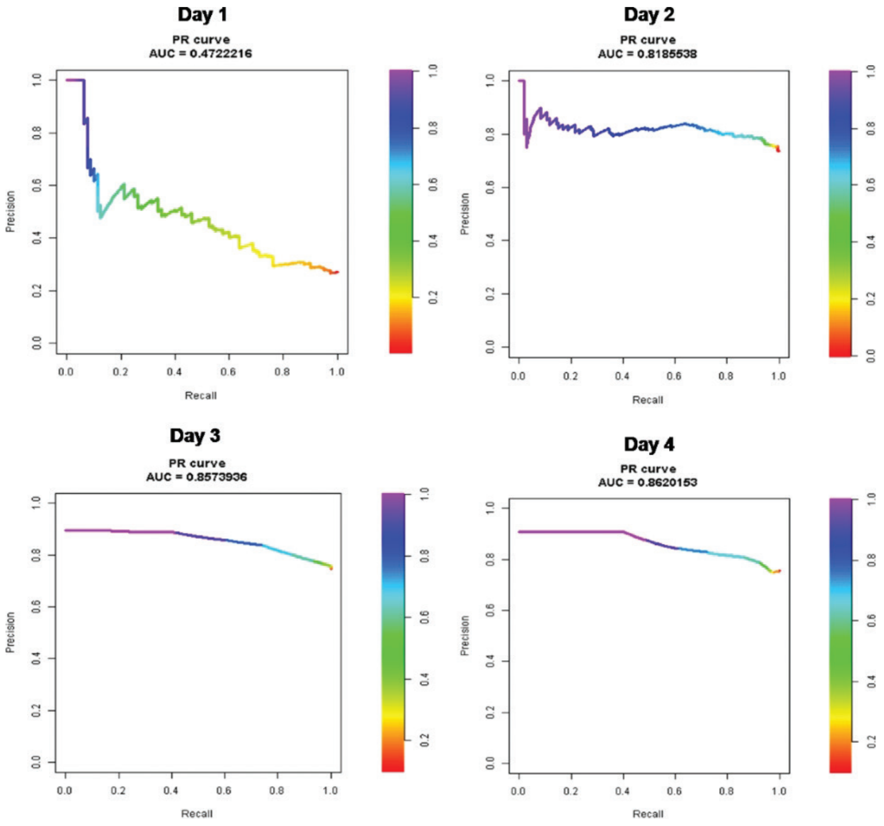

Figure 2. PR curves for first 4 days that were generated from the predictions of the best performing models, built on features from the Gini selection.

Day	J4.8	Logistic	Random Forest	AdaBoost (J 4.8)	AdaBoost (Logistic)	Bagging (J4.8)	Bagging (Logistic)
1	0.35 (10)	**0.47 (60)**	0.42 (30)	0.37 (10)	0.42 (30)	0.36 (20)	0.46 (30)
2	0.76 (60)	**0.83 (50)**	**0.83 (70)**	0.81 (30)	0.82 (60)	0.78 (50)	0.82 (60)
3	0.78 (30)	0.84 (20)	**0.85 (40)**	0.84 (60)	0.82 (90)	0.76 (10)	0.83 (20)
4	0.79 (20)	0.85 (20)	**0.86 (40)**	**0.86 (10)**	0.83 (20)	0.78 (10)	0.84 (10)
5	0.81 (20)	0.85 (30)	**0.86 (80)**	0.85 (20)	0.83 (20)	0.8 (10)	0.84 (10)
6	0.82 (20)	0.83 (40)	**0.86 (70)**	**0.86 (20)**	0.82 (90)	0.8 (20)	0.83 (10)
7	0.80 (10)	0.84 (30)	**0.86 (80)**	0.85 (10)	0.81 (30)	0.78 (10)	0.84 (30)

Bold values represent the best results per rows. Multiple bold Values per row means that there was more than one equally good results.

Table 3. Algorithms performance per admission day based on Gini index feature selection measure.

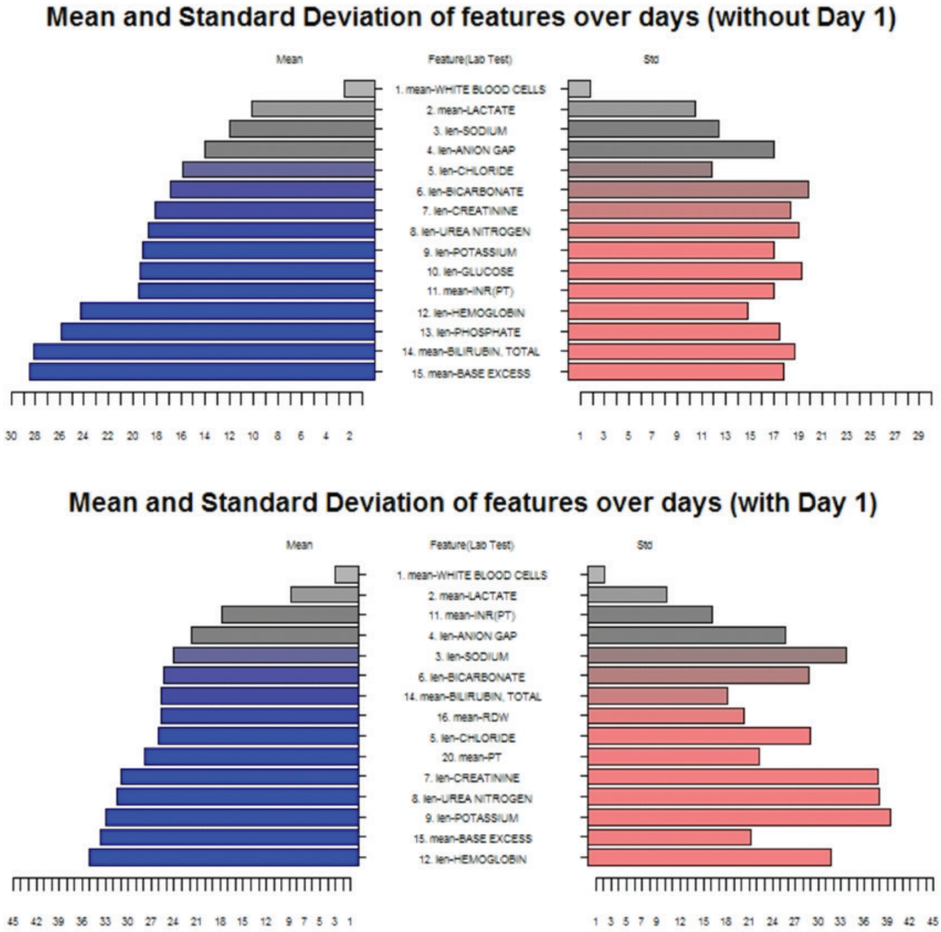

Figure 3. Pyramid chart of feature ranking (laboratory tests) for predictive value by mean and standard deviation; in- and -excluding day 1.

Finally, we analyzed if there was the difference, between features selected on the first day, where the predictive performance was consistently poor, and other days, where the predictive performance was acceptable. Rank means and standard deviations were calculated for two groups: all days (with day one) and days 2–7 (without day 1) (**Figure 3**). Standard deviations of ranks are much higher over ranks of features that include day 1 (right parts of **Figure 3**). The average ranks changed (middle part of the figure), but similar laboratory tests were in first 15 ranks in both cases.

4. Discussion

This study using ensemble methods demonstrated an improvement in predictive accuracy compared to prediction based on single models. Random Forests seem to provide the best predictive accuracy complying with our previous research [8]. (**Table 3**) Random Forest also resulted in a high predictive accuracy for mortality risk prediction [29]. This study, however, did not analyze different ensemble and feature selection methods and was conducted on a different population. In addition, the feature selection techniques Gini Index and ReliefF scored best in the majority of the cases.

Laboratory tests ranked per day based on Random Forest and Gini Index. (*mean*, *std.*: standard deviation, *len*: number of tests/day) indicated the importance of mean lactate values (ranked first on day 1, 2, and 8), and mean white blood cells count (on day 3, 4 and 6) in the prediction of hospital mortality. In addition, it is fascinating to observe that predictive features for hospital mortality calculated from 755 laboratory related parameters and without any additional patient-related information or medical knowledge, correlated well with the laboratory tests used on a daily basis (sodium, anion gap, chloride, bicarbonate, creatinine, urea nitrogen, potassium, glucose, INR, hemoglobin, phosphate, total bilirubin, and base excess). Laboratory-based clinical decision support may improve physician adherence to guidelines with respect to timely monitoring of chronic kidney disease [30, 31].

As demonstrated in this study, parameters for shock (lactate), sepsis (white blood cells), and multi-organ failure are more important [32].

The ensemble models in this study were able to generate a high predictive accuracy (AUPRC values) from day 4, with acceptable results on the second and third day. On the first day of admission, however, AUPRC values were very low but correlated well with diagnostic uncertainty on the first day of admission.

Surprisingly, patients who survived hospital stay were significantly older but suffered less from lymphoma and metastatic cancer. These findings might indicate some admission bias for certain comorbidities or indicate a constitutional superiority of older people admitted to ICU despite renal failure.

The hospital mortality in this renal failure ICU population was 27.0% (399/1477) [33]. Laboratory testing alone is only a part of the daily assessment of ICU patients. More research could elaborate predictive analysis of laboratory tests and other patient-related data in different patient populations.

5. Conclusions

Predictive analytics using ensemble methods are able to predict hospital or ICU outcome of renal patients with high accuracy. Predictive accuracy changes with the length of stay. Feature ranking enables quantitative assessment of patient data (e.g. laboratory tests) for predictive power. Lactate and white blood cell count best predict hospital mortality in this population. From the second day of ICU admission, predictive accuracy based on laboratory tests >80%. This generates opportunities for efficacy and efficiency analysis of other data recorded during ICU stay.

Acknowledgements

Licenses for RapidMiner/Radoop are provided by RapidMiner (Cambridge, MA, and USA).

Support for data extraction was kindly provided by Wessel Luijben from Anchormen (Amsterdam, The Netherlands). RapidMiner (RapidMinerAcademia) provided a free use of the commercial version of its platform to researchers and other academics at educational institutions.

Author details

Sven Van Poucke[1], Ana Kovacevic[2] and Milan Vukicevic[3]*

*Address all correspondence to: vukicevicm@fon.bg.ac.rs

1 Department of Anesthesiology, Intensive Care, Emergency Medicine and Pain Therapy, Ziekenhuis Oost-Limburg, Genk, Belgium

2 Saga Ltd., New Frontier Group, Belgrade, Serbia

3 Faculty of Organizational Sciences, University of Belgrade, Belgrade, Serbia

References

[1] Yan Q. From pharmacogenomics and systems biology to personalized care: A framework of systems and dynamical medicine. Methods in Molecular Biology. 2014;**1175**:3-17

[2] Ullman AJ, Keogh S, Coyer F, et al. "True Blood" The Critical Care Story: An Audit of Blood Sampling Practice Across Three Adult, Paediatric and Neonatal Intensive Care Settings [Internet]. Australian Critical Care. 2015. Available from: http://www.sciencedirect.com/science/article/pii/S1036731415000752

[3] Ezzie ME, Aberegg SK, O'Brien JM. Laboratory testing in the intensive care unit. [Internet]. Critical Care Clinics. 2007;**23**:435-465

[4] Frassica JJ. Frequency of laboratory test utilization in the intensive care unit and its implications for large-scale data collection efforts. Journal of the American Medical Informatics Association. 2005;**12**:229-233

[5] Yurkovich M, Avina-Zubieta JA, Thomas J, et al. A systematic review identifies valid comorbidity indices derived from administrative health data. Journal of Clinical Epidemiology. 2015;**68**:3-14

[6] Garvin JH, Redd A, Bolton D, et al. Exploration of ICD-9-CM coding of chronic disease within the Elixhauser comorbidity measure in patients with chronic heart failure. Perspectives in Health Information Management. 2013;**10**(Fall):1b

[7] Southern DA, Quan H, Ghali WA. Comparison of the Elixhauser and Charlson/Deyo methods of comorbidity measurement in administrative data. Medical Care. 2004;**42**(4): 355-360

[8] Van Poucke S, Zhang Z, Schmitz M, Vukicevic M, Vander Laenen M, Celi LA, De Deyne C. Scalable predictive analysis in critically ill patients using a visual open data analysis platform. PLoS One. 2016;**11**(1)

[9] Ritthoff O, Klinkenberg R, Fisher S, Mierswa I, Felske S. YALE: Yet Another Learning Environment. LLWA'01–Tagungsband der GI-Workshop-Woche Lernen–Lehren–Wissen Adaptivitat. Dortmund, Germany: University of Dortmund. Technical Report 763. 2001. pp. 84-92

[10] Zhang Z. Data management by using R: Big data clinical research series. Annals of Translational Medicine. 2015;**3**(20):303. DOI: 10.3978/j.issn.2305-5839.2015.11.26

[11] Zhang Z. Missing values in big data research: Some basic skills. Annals of Translational Medicine. 2015;**3**(21):323. DOI: 10.3978/j.issn.2305-5839.2015.12.11

[12] Goldberger AL, Amaral LAN, Glass L, et al. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation. 2000;**101**:215-220

[13] Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): A public-access intensive care unit database. Critical Care Medicine. 2011;**39**(5):952-960. DOI: 10.1097/CCM.0b013e31820a92c6

[14] McGregor C, Catley C, James A. A process mining driven framework for clinical guideline improvement in critical care. In: CEUR Workshop Proceedings; 2011

[15] Chao C-M, Yu Y-W, Cheng B-W, et al. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. Journal of Medical Systems. 2014;**38**:106

[16] Ting H, Mai Y-T, Hsu H-C, et al. Decision tree based diagnostic system for moderate to severe obstructive sleep apnea. Journal of Medical Systems. 2014;**38**:94

[17] Quinlan JR. Induction of decision trees. Machine Learning. 1986;**1**:81-106

[18] Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: Data mining, inference and prediction. Mathematical Intelligence. 2005;**27**:83-85

[19] Druss BG, Marcus SC, Rosenheck RA, Olfson M, Tanielian T, Pincus HA. Understanding disability in mental and general medical conditions. The American Journal of Psychiatry. 2000;**157**(9):1485-1491

[20] Post RM, Altshuler L, Leverich GS, Frye MA, Suppes T, McElroy SL, et al. Relationship of clinical course of illness variables to medical comorbidities in 900 adult outpatients with bipolar disorder. Comprehensive Psychiatry. 2015;**56**:21-28

[21] Kuncheva LI, Whitaker CJ. Measures of diversity in classifier ensembles. Machine Learning. 2003;**51**:181-207

[22] Breiman L. Random forests. Machine Learning. 2001;**45**:5-32

[23] Freund Y, Schapire R, Abe N. A short introduction to boosting. Journal of JSAI. 1999;**14**(5):771-780

[24] Breiman L. Bagging predictors. Machine Learning. 1999;**24**(2):123-140

[25] Kononenko I. Estimating attributes: Analysis and extensions of RELIEF. In: European Conference on Machine Learning. Berlin, Heidelberg: Springer; 1994. pp. 171-182

[26] Shearer C. The CRISP-DM model: The new blueprint for data mining. Journal of Data Warehousing. 2000;**5**:13-22

[27] Riley RD, Ahmed I, Debray TPA, Willis BH, Noordzij JP, Higgins JPT, Deeks J. Summarising and validating test accuracy results across multiple studies for use in clinical practice. Statistics in Medicine. 2015;**34**(13):1097-0258

[28] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine learning; (ICML 2006). New York, NY, USA: ACM; pp. 233-240

[29] Churpek MM, Yuen TC, Winslow C, Meltzer DO, Kattan MW, Edelson DP. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. Critical Care Medicine. 2016;**44**(2):368-374

[30] Rassa AC, Horne BD, McCubrey RO, Bair TL, Muhlestein JB, Morris DR, Anderson JL. Novel stratification of mortality risk by kidney disease stage. American Journal of Nephrology. 2015;**42**(6):443-450

[31] Ennis J, Gillen D, Rubenstein A, et al. Clinical decision support improves physician guideline adherence for laboratory monitoring of chronic kidney disease: A matched cohort study. BMC Nephrology. 2015;**16**:163. DOI: 10.1186/s12882-015-0159-5

[32] Zhang Z, Ni H. Normalized lactate load is associated with development of acute kidney injury in patients who underwent cardiopulmonary bypass surgery. In: Ricci Z, editor. PLoS One. 2015;**10**(3):e0120466

[33] Timmers TK, Verhofstad MH, Moons KG, et al. Long-term survival after surgical intensive care unit admission: Fifty percent die within 10 years. Annals of Surgery. 2011;**253**: 151-157