Rating organ failure via adverse events using data mining in the intensive care unit

Álvaro Silva[a], Paulo Cortez[b⋆], Manuel Filipe Santos[b],

Lopes Gomes[c], José Neves[d]

[a] Serviço de Cuidados Intensivos, Hospital Geral de Santo António, Porto, Portugal

[b] Departamento de Sistemas de Informação, Universidade do Minho, Guimarães, PORTUGAL

[c] Clínica Médica I, Inst. de Ciências Biomédicas Abel Salazar, Porto, Portugal

[d] Departamento de Informática, Universidade do Minho, Braga, PORTUGAL

⋆ Corresponding author:

Paulo Cortez

Tel: +351-253-510313; fax: +351-253-510300.

E-mail: pcortez@dsi.uminho.pt

Departamento de Sistemas de Informação, Universidade do Minho, Campus de Azurém, 4800-058 Guimarães, PORTUGAL

Summary

Objective: The main intensive care unit (ICU) goal is to avoid or reverse the organ failure process by adopting a timely intervention. Within this context, early identification of organ impairment is a key issue. The sequential organ failure assessment (SOFA) is an expert-driven score that is widely used in European ICUs to quantify organ disorder. This work proposes a complementary data-driven approach based on adverse events, defined from commonly monitored biometrics. The aim is to study the impact of these events when predicting the risk of ICU organ failure.

Materials and Methods: A large database was considered, with a total of 25215 daily records taken from 4425 patients and forty two European ICUs. The input variables include the case mix (i.e. age, diagnosis, admission type and admission from) and adverse events defined from four bedside physiologic variables (i.e. systolic blood pressure, heart rate, pulse oximeter oxygen saturation and urine output). The output target is the organ status (i.e. normal, dysfunction or failure) of six organ systems (respiratory, coagulation, hepatic, cardiovascular, neurological and renal), as measured by the SOFA score. Two data mining (DM) methods were compared: multinomial logistic regression (MLR) and artificial neural networks (ANNs). These methods were tested in the R statistical environment, using twenty runs of a 5-fold cross-validation scheme. The area under the receiver operator characteristic (ROC) curve and Brier score were used as the discrimination and calibration measures.

Results: The best performance was obtained by the ANNs, outperforming the MLR in both discrimination and calibration criteria. The ANNs obtained an average (over all organs) area under the ROC curve of 64%, 69% and 74% and Brier scores of 0.18, 0.16 and 0.09 for the dysfunction, normal and failure organ conditions respectively. In particular, very good results were achieved when predicting renal failure (ROC curve area of 76% and Brier Score of 0.06).

Conclusion: Adverse events, taken from bedside monitored data, are important intermediate outcomes, contributing to a timely recognition of organ dysfunction and failure during ICU length of stay. The obtained results show that is possible to use DM methods to get knowledge from easy obtainable data, thus opening room for the development of intelligent clinical alarm monitoring.

Keywords: Adverse event; Artificial neural networks; Critical care; Data mining; Multinomial logistic regression; Organ failure assessment.

1 Introduction

Since the early 1980s clinical scores have been developed to access severity of illness and organ dysfunction in the intensive care unit (ICU) setting [1]. Indeed, in the context of intensive medicine, severity scores are instruments that aim primarily at stratifying patients based on risk adjustment of the clinical condition. Furthermore, these tools have been used to improve the quality of intensive care and guide local planning of resources.

The majority of these scores use are static, since they use data collected only on the first ICU day, such as as the acute physiology and chronic health evaluation system (APACHE) [2], the simplified acute physiology score (SAPS) [3] or mortality probability model (MPM) [4]. Yet, these static scores fail to recognize several factors that can influence the patient outcome after the first 24 hours (e.g. the therapeutics strategy and the patients' response).

More recently, dynamic (or repetitive) scores have been designed, where the data and scores are updated on a daily basis. The most used scores include [5]: the sequential organ failure assessment (SOFA), multiple organs dysfunction score (MODS) and logistic organ dysfunction (LOD). Our focus is on the SOFA score which was first proposed to evaluate morbidity (degree of organ failure) [6] and latter it has been shown to be related with mortality risk [7, 8].

The SOFA scores six organ systems (respiratory, coagulation, hepatic, cardiovascular, neurological and renal) on a scale ranging from 0 to 4, according to the degree of failure. This is an expert-driven score, in the sense that it was developed by a panel of experts who choose a set of variables and rules based on their personal opinions [5]. The SOFA is widely used in European ICUs, nevertheless there are some issues not yet solved. Firstly, for some of the variables (e.g. platelets and bilirubin), the SOFA uses the worst value obtained in the last 24 hours and it is not clear how many daily times they should be measured. Also, the SOFA is a classification system that does not provide a risk (i.e. probability) of the outcome of interest (i.e. organ failure).

On the other hand, bedside monitoring of physiologic variables is universal and routinely registered during patient ICU stay. Indeed, ICU physicians tend to analyze these monitoring data in an empirical fashion in order to trigger an action given a specific condition. The relationships within these data are complex, nonlinear and not fully understood. For instance, if a severe arterial hypotension (i.e. low blood pressure) arises then renal or cardiovascular failure may succeed. Yet, it is not clear what should be the duration and/or severity of the hypotension to trigger the latter outcomes. Thus, monitoring analysis is not standardized and mainly relies on the physicians knowledge and experience to interpret them. The SOFA score uses both physiological parameters (e.g. hypotension) and laboratory data (e.g platelets). However, the latter ones usually depend on previous physiological impairments. For example, a severe and long hypotension associated with hypoxemia can lead to hepatic failure (i.e. bilirubin increase). Therefore, using only biometric data should potentially allow a more adequate evaluation and early therapeutic intervention.

Yet, as more and more biometrics are continuously monitored (e.g. mechanical ventilator, cardiovascular device), the amount of data available increases exponentially, generating alarms that need to be interpreted. In previous work [9], we have shown that out of range measurements (or adverse events) of four biometrics (i.e. systolic blood pressure, heart rate, pulse oximeter oxygen saturation and urine output) have an impact on the mortality outcome of ICU patients. Since multiple organ failure is a major cause for ICU mortality [8], it is rational to access the impact of the adverse events on organ system function at an early stage.

One of the most promising recent developments in intensive care consists in the use of artificial intelligence/data mining techniques [1, 10]. The fast growing amount of data collected had led to vast and complex databases that exceeded the human capability for comprehension without using computational resources. The goal of data mining (DM) is to discover interesting knowledge from the raw data by using automatic discovery tools [11].

There are several DM techniques, each one with its own purposes and advantages. The majority of the severity scores use statistical methods such as the logistic regression (LR), which is easy to interpret. Yet, such classical statistics may not be suitable for the complex nonlinear relationships often found in biomedical data [1]. Artificial neural networks (ANNs) are connectionist models inspired by the behavior of the human brain [12]. In ICUs, ANNs are gaining an increase of acceptance due to advantages of nonlinear learning and high flexibility. Indeed, ANNs have been applied to predict mortality and length of stay [1, 10].

Motivated by the results obtained in [13], a novel approach is presented in this work, where the main goal is to explore the impact of the adverse events, during the last 24h, on the current day organ risk condition (i.e. normal, dysfunction or failure). As a secondary goal, two DM techniques (i.e. LR and ANNs) are evaluated and compared. The proposed approach will be tested on a large database, which includes daily records of 4425 patients taken from forty two European ICUs.

The paper is organized as follows. Section 2 presents the ICU clinical data, DM models, feature selection approach and computational environment. Next, the results are analyzed (Section 3) and discussed (Section 4). Finally, closing conclusions are drawn (Section 5).

2 Materials and methods

2.1 Intensive care data

The database used in the present study was constructed by the authors from the EURICUS II study. The EURICUS II project was conducted from November/98 to August/99 and encompassed forty two ICUs from nine European Union countries (see [14] for more details).

In each participating ICU, monitoring data was collected and registered manually. According to the universal monitoring practice, in every hour, all ICU patient biometrics were recorded in a standardized sheet form by the nursing staff. Also, the adverse events were assigned in a specific sheet at a hourly basis. The registered data was submitted to a double check, using both local (i.e. ICU) and central levels (i.e. Health Services Research Unit of the Groningen University Hospital, the Netherlands). The latter unit was used to gather the full database.

Two main criteria were used for the event definition. First, its occurrence and duration should be registered by physiological changes (e.g. shock and not pneumonia). Second, the related physiological variables should be routinely registered at regular intervals. Four biometrics filled these requirements: the systolic blood pressure (BP), the heart rate (HR), the pulse oximeter oxygen saturation (SpO_2) and the hourly urine output (UR). The normal ranges for these parameters (see Table 1) were set by a panel of seven experts. An alarm is triggered if there is an out of range value during a given time, defining an event. It should be noted that the minimum time period was set to $10min$ to minimize the number of false alarms triggered by technical problems (e.g. disconnected sensor). For each biometric, the daily number of events were stored. When a longer event occurs or a more extreme physiologic measurement is found, it is called a critical event. For this last case, the database includes daily entries with the number of critical events and its duration. Table 2 shows a synopsis of the ICU variables considered. The first four attributes (the case mix) are static, being collected during the patient's admission. The next twelve variables are related to the adverse events.

At a daily basis, the SOFA score was computed for six organ systems (respiratory, coagulation, hepatic, cardiovascular, neurological and renal) by collecting the raw data presented in Table 3 during the last 24h. The SOFA values range from 0 to 4, with the following interpretation: 0 – normal; 1 or 2 – dysfunction; 3 or 4 – failure.

*** insert Table 1 around here ***

*** insert Table 2 around here ***

The exclusion criteria fulfilled the SAPSII definitions [3], i.e. with age lower than eighteen years old, burned or with recent coronary bypass surgery. Also, the last day of stay data entries were discarded, since the SOFA score is only defined for a 24h time frame and several of these patients were discharged earlier. The final database contains a total of 25215 daily records taken from 4425 critically ill patients.

Figure 1 plots the histograms of the SOFA values for each organ (computed over the whole database). The figure shows the prevalence of each condition, denoting skewed distributions, i.e. the number of normal conditions is higher than the failure ones. During the preprocessing stage, each SOFA variable was transformed into a three-class output, one for each organ condition: normal, dysfunction and failure.

*** insert Table 3 around here ***

*** insert Figure 1 around here ***

For demonstrative purposes, Figure 2 presents the boxplots of the time of critical events associated to each renal status. In the boxplots, it is difficult to find a clear pattern that relates adverse events to the organ condition, suggesting that this is a non trivial task.

*** insert Figure 2 around here ***

2.2 Data mining methods

Data mining (DM) is an emerging area that lies at the intersection of statistics, artificial intelligence and data management. DM tasks can be classified into two categories [11]: descriptive, where the intention is to characterize the properties of the data; and predictive, to forecast the unknown value of an output target given known values of other variables (the inputs). Predictive tasks can be further divided into classification, when the output domain is discrete, and regression, when the dependent variable is continuous.

The multinomial logistic regression (MLR) is the extension of the common logistic method to multi-class tasks. Let $c_j \in C$ be the condition j and C the set of all possible classes, then the respective estimated probability ($\widehat{p}_j$) is given by [15]:

$$\begin{aligned} \widehat{p}_j &= \frac{exp(\eta_j \mathbf{x})}{\sum_{k=1}^{\#C} exp(\eta_k \mathbf{x})} \\ \eta_j(\mathbf{x}) &= \sum_{i=1}^{I} \beta_{j,i} x_i \end{aligned} \tag{1}$$

where $\beta_{j,0}, \ldots, \beta_{j,I}$ denotes the parameters of the model, and $x_1, \ldots, x_I$ the dependent variables. This model requires that $\eta_k(\mathbf{x}) \equiv 0$ for one $c_k \in C$ (the baseline group) and this assures that $\sum_{j=1}^{\#C} \widehat{p}_j = 1$. It should be noted that the selection of the baseline class (c_k) does not affect the MLR performance.

The multilayer perceptron is a popular artificial neural network (ANN), where processing neurons are grouped into layers and connected by weighted links [12]. The ANN is activated by feeding the input layer with the input variables and then propagating the activations in a feedforward fashion, via the weighted connections, through the entire network.

A fully connected network, with one hidden layer of H nodes, will be adopted in this work. For multi-class data, the ANN outputs can be interpreted as probabilities if the logistic function is applied to the hidden neurons and the linear function is used at the $\#C$ output nodes. Then, the final ANN probability estimate for the class j is given by [15]:

$$\begin{array}{rcll} \widehat{p}_j & = & \frac{exp(y_j)}{\sum_{k=1}^{\#C} exp(y_k)} & \text{(softmax function)} \\ y_i & = & w_{i,0} + \sum_{m=I+1}^{I+H} f(\sum_{n=1}^{I} x_n w_{m,n} + w_{m,0}) w_{i,n} & \end{array} \qquad (2)$$

where y_i is the output of the network for the node i; $f = \frac{1}{1+exp(-x)}$ is the logistic function; I represents the number of input neurons; $w_{d,s}$ the weight of the connection between nodes s and d; and $w_{d,0}$ is a constant called bias. The first equation, known as the *softmax* function, warranties that $\widehat{p}_j \in [0,1]$ and $\sum_{j=1}^{\#C} \widehat{p}_j = 1$. The simplest ANN (with $H = 0$) is equivalent to the MLR model and more complex discrimination functions can be learned with a higher number of hidden neurons (Figure 3). Yet, a high value of H will induce generalization loss (i.e. overfitting).

The logistic model is easier to interpret than ANNs. Nevertheless, it is possible to gather knowledge about what the ANN has learned by measuring the relative importance of the inputs (Section 2.3) and extracting rules. The latter issue is still an active research domain [16]. In this work, the pedagogical technique presented in [9] will be adopted, where the direct relationships between the inputs and outputs of the ANN are extracted by using a decision tree [17].

*** insert Figure 3 around here ***

2.3 Sensitivity analysis and feature selection

The sensitivity analysis [18] is a simple procedure that analyses the model responses when the inputs are changed. Although originally proposed for ANNs, this sensitivity method can also be applied to other DM models, such as logistic regression or support vector machines [19]. Let $\widehat{p}_{c_j}^{i}$ denote the probability of condition c_j when all input variables are hold at their average values. The exception is the attribute x_a, which varies through its range with $i \in \{1, \ldots, L\}$ levels. In this work, we will adopt the average gradient (G_a) as the sensitivity measure. For a multi-class domain, it is given by:

$$\begin{array}{rcl} G_a & = & \frac{\sum_{j=1}^{\#C} \sum_{i=1}^{L-1} |\widehat{p}_{c_j}^{i+1} - \widehat{p}_{c_j}^{i}|}{\#C(L-1)} \\ R_a & = & V_a / \sum_{k=1}^{A} G_k \end{array} \qquad (3)$$

where A denotes the number of input attributes and R_a the relative importance of attribute a (in %). In the experiments, L will be set to the number of discrete values for the nominal attributes and 6 for the continuous inputs ($x_a \in \{-1.0, -0.6, \ldots, 1.0\}$).

Feature selection methods [20] are useful to discard irrelevant inputs, leading to simpler models that are easier to interpret and often presenting higher predictive accuracies. A covariance analysis was applied to the attributes of Table 2, revealing weak relationships except for the variables related to the same biometric (e.g. the correlation between NCRBP and TCRBP is 0.7). This suggests that the number of irrelevant features is low, although the covariance procedure is only capable of measuring linear dependences. Therefore, a backward variable selection method will be applied to both the MLR and ANN models.

The backward search will be guided by the sensitivity measure [18], allowing a reduction of the computational effort by a factor of A when compared to the standard backward selection algorithm [20]. All inputs are used at the beginning and the data is randomly split into training (66.6%) and validation (33.3%) sets. In each iteration, the former set is used to fit the model and get the importance values (R_a), while the validation data is used to access the generalization error. Then, the least relevant feature (i.e. with the lowest R_a) is discarded. The process is repeated until there is no error improvement during E iterations (in this work set to $E = 3$) or after A cycles. Finally, the lowest validation error is the criterion for selecting the best set of variables.

2.4 Evaluation

The receiver operating characteristic (ROC) curve shows the performance of a two class classifier across the range of possible threshold (D) values, plotting one minus the specificity (x-axis) versus the sensitivity (y-axis) [21]. The overall accuracy is given by the area under the curve ($AUC = \int_0^1 ROC dD$), measuring the degree of discrimination that can be obtained from a given model. In intensive care, the AUC is the most popular metric for prognostic scores [10], where the ideal method should present an AUC of 1.0, while an AUC of 0.5 denotes a random classifier. In the medical literature, values of AUC above 0.7 are considered acceptable [1, 10]. Multi-class problems can be handled by producing one ROC for each class [21]. The ROC graph for the class reference c_i is generated by considering the positive (c_i) and negative ($C \setminus c_i$) labels. The global AUC can then be computed by summing the AUCs weighted by the prevalence of c_i in the data, using [22]:

$$\begin{aligned} AUC_{Global} &= \sum_{c_i \in C} AUC(c_i) \cdot prev(c_i) \\ prev(c_i) &= \#c_i / N \end{aligned} \tag{4}$$

where $AUC(c_i)$ denotes the AUC for class reference c_i, $\#c_i$ the number of patients with condition c_i and N the total number of patients.

Another important criterion is the calibration, which measures how close the predictions ($\widehat{p}$) are to the true probabilities (p) of an event. In this work, calibration will be assessed using the widely used Brier score ($\in [0, 1]$), which is defined for a two-class scenario as [23]:

$$Brier(c_j) = \frac{1}{N}\sum_{i=1}^{N}(p_j^i - \widehat{p}_j^i)^2 \tag{5}$$

where p_j^i and $\widehat{p}_j^i$ denote the actual c_j outcome (0 or 1) for the patient i and respective probability estimation. Inspired in the multi-class AUC metric, the global Brier score is defined as:

$$Brier_{Global} = \sum_{c_i \in C} Brier(c_i) \cdot prev(c_i) \tag{6}$$

The lower the value, the better is the calibration, with the perfect model presenting a Brier score of 0.

Calibration can also be visualized with the regression error characteristic (REC) curve [24], which is used to compare regression models and it plots the error tolerance (x-axis), given in terms of the absolute deviation, versus the percentage of points predicted within the tolerance (y-axis). Similarly to the ROC concept, the ideal regressor should present a REC area of 1.0.

The $K-$fold cross-validation [25] is a commonly used method to estimate generalization performances. In each run, the data is divided into K partitions of equal size. Sequentially, one different subset is tested and the remaining data is used for fitting the model. Under this scheme, all data is used for testing, although K different models are fitted. This work will use 20 runs of a 5-fold, in a total of 20×5=100 experiments for each tested configuration. Statistical significance for the AUC and

Brier values will be given by using a Mann-Whitney non-parametric test at the 95% confidence level. According to [26], this test is equivalent to the test proposed by DeLong et al. [27] to compare ROC areas.

2.5 Computational environment

All experiments were conducted using the **RMiner** [28], an open source library for the **R** statistical environment [29] that facilitates the use of DM techniques in classification and regression tasks. In particular, the **RMiner** uses the `multinomial` and `nnet` functions of the **nnet** package to implement the MLR and ANN models [15]. Also, the efficient Algorithms 1 and 2 presented in [21] are used to compute the ROC curves and AUC values.

In this work, we will adopt the default suggestions of the **nnet** developers [15] to adjust the DM techniques. The nominal inputs were encoded into *1-of-(*$\#C-1$*)* binary variables. As an example, **admtype** from Table 2 is transformed with: $1 \rightarrow (0\ 0)$; $2 \rightarrow (1\ 0)$; and $3 \rightarrow (0\ 1)$. For the ANNs, the continuous inputs were scaled into a zero mean and one standard deviation range. Both the MLR and ANN models were trained using 100 iterations (known as epochs) of the efficient BFGS algorithm [30], from the family of quasi-Newton methods. Within a given epoch, the whole training dataset is presented to the ANN, in order to compute an error function that is used to adjust the neural weights. For multi-class data, the algorithm is set to maximize the likelihood, which is equivalent to minimizing the cost error function (ξ) given by:

$$\xi = \sum_{i=1}^{N} \sum_{j=1}^{\#C} [p_j^i ln \frac{p_j^i}{\widehat{p}_j^i} + (1-p_j^i) ln \frac{1-p_j^i}{1-\widehat{p}_j^i}] \qquad (7)$$

In contrast with the MLR, the adopted ANN model requires the definition of one hyperparameter, the number of hidden nodes (H). To set this value, the **RMiner** provides a grid search facility, where $H \in \{H_L, H_L + g, H_L + 2g, \ldots, H_U\}$, H_L and H_U denote the lower and upper bounds; and g is a constant value. To prevent the overfitting phenomenon and also to reduce the search time, we will adopt a small range (i.e. $H \in \{2, 4, 6, 8, 10\}$). Also, and due to computational limitations, H will be fixed to the median of the grid range during the feature selection phase [19]. Then, the grid search is applied, using a random $\frac{2}{3}/\frac{1}{3}$ data split for the training and validation sets. The best H will be the one that provides the lowest validation error. After selecting the best attributes and H value (in case of ANN), the final model is retrained with all available data.

3 Results

3.1 Predictive performance

A total of 6 (organs) $\times$ 2 (methods) = 12 different configurations were tested. The median number of the selected hidden nodes was 8 for all organs except the neurological, where the median was 10. For tested configurations, the feature selection algorithm only discarded an average of 2 attributes. In general, the few removed variables are related to the adverse events. Nevertheless, all four biometrics are used in all models (e.g. NCRUR may be deleted but TCRUR is not). These results confirm the covariance analysis performed on Section 2.3.

*** insert Table 4 around here ***

The discrimination results evaluated over the test sets are summarized in Table 4. The best results are obtained by the ANNs, which outperform the MLR with an average (last row) margin of 2.2, 1.8 and 2.8 percentage points for the normal, dysfunction and failure status respectively. The AUC differences (ANN vs MLR) are significant (p-value$<$ 0.05) in all cases. When analyzing the organ condition discrimination, the dysfunction condition is more difficult to predict. In effect, none of the presented models has acceptable values (AUC higher than 70%). The normal status shows a higher discrimination, with 1 MLR and 3 ANN acceptable models. Finally, the failure condition presents the most accurate predictions. The MLR models are acceptable for the coagulation, hepatic, neurological and renal systems, while the ANNs obtain good performances for all organs except respiratory. In particular, the hepatic, neurological and renal AUCs are above 75%. When weighted by the condition prevalence, the global AUC reveals three acceptable models (ANN for the cardiovascular, neurological and renal systems). All ROC curves are plotted in Figure 4. In the graphs, the ANN curves are above the MLR ones, confirming the superiority of the discrimination power of the ANNs.

The calibration results are presented in Table 5. The global Brier scores are particularly good for both DM methods on three organs (coagulation, hepatic and renal). Nonetheless, the ANN outperforms the logistic model in all cases except the hepatic dysfunction and coagulation failure conditions (the differences are significant, with p-value$<$ 0.05). Regarding the organ status, the best calibration is obtained for the failure state (average Brier score for all organs of 0.093), followed by the dysfunction (0.156) and normal (0.181) conditions. These results are complemented by a REC analysis (Figure 5). High quality curves (REC close to 1) were achieved for the prediction of the coagulation, hepatic and renal failures, precisely where lowest Brier scores were obtained. Although MLR and ANN curves are close, latter ones present a higher area. Also, more patient conditions are correctly predicted for low admitted errors. For instance, if a 0.1 tolerance is accepted (e.g. a

0.9 output is interpreted as positive), then 27.7% of the coagulation failure (positive or negative) examples are correctly estimated for the ANN method. This value decreases to 18% for the MLR.

3.2 Descriptive knowledge

This section will provide explanatory knowledge that can be useful for the intensive care domain. The goal is not to infer about the predictive capabilities of each model, as measured in the Section 3.1, but to give a simple description that summarizes the DM models. Thus, the whole dataset will be used in the descriptive experiments.

Tables 6 and 7 present the relevance (in percentage) of each input variable for the two DM methods. For both MLR and ANN, the four biometrics are important for all organs, although the relative impact may differ. For the logistic model, the adverse events overall influence ranges from 52.5% (cardiovascular) to 69.8% (hepatic), while the interval varies from 38.6% (coagulation) to 50.3% (respiration) for the ANN. Regarding the MLR model, the most important biometrics are on average the oxygen saturation and heart rate. The oxygen alarms are also the most relevant for the ANNs, followed by the blood pressure.

For demonstrative purposes, more detail will be given to the renal models, which obtained satisfactory discrimination and calibration values. Table 8 shows the β_i, j MLR coefficients (the model was fitted with all available data). The R environment automatically selected the dysfunction class as the baseline group, thus $\widehat{p}_{dysfunction} = 1 - (\widehat{p}_{failure} + \widehat{p}_{normal})$ and no coefficients are used by this condition. These coefficients should not be read separately, since organ function condition results from the impact of complex interactions between all physiological metrics. For instance, regarding the urine output, while the values suggest that renal failure is negatively influenced by the number of events (NUR), it is also positively influenced by long lasting critical events ($TCRUR$).

In this example, the feature selection algorithm discarded one variable (NCRUR) for the MLR, while the final neural model did not include 3 attributes (NCRSpO_2, TCRBP, TCRHR). The latter contains 19 input, 8 hidden and 3 output neurons, with a total of 187 weights. Instead of presenting all these weights, and to simplify the analysis, a decision tree will be used to describe the ANN behavior [9]. The tree was fit using the default values of the **rpart R** library [15] and a training set composed by the ANN inputs and outputs. The latter ones were preprocessed into the condition related to the highest ANN probability. The obtained model (Figure 6) managed to mimic the ANN behavior with a low classification error (3.4%) and it includes the two most relevant biometrics from Table 7 (UR ad HR). As an example, the next two rules for renal failure prediction can be extracted from the tree:

$$\begin{aligned}
&\text{IF TCRUR} \geq 13.8 \text{ AND NUR} \geq 15 \text{ THEN } \textit{failure} \\
&\text{IF TCRUR} < 13.8 \text{ AND admfrom} \notin \{5,6\} \\
&\quad \text{AND NCRHR} = 0 \text{ AND SAPSII} \geq 93 \text{ THEN } \textit{failure}
\end{aligned} \tag{8}$$

4 Discussion

The assessment of the degree of organ failure is crucial in intensive care units (ICUs), since one of the main ICU tasks is to avoid or reverse organ failure process by an early identification of patients at risk and adopting the respective therapy. Indeed, several expert-driven scores have been developed to quantify organ disorder, such as the sequential organ failure assessment (SOFA), which is widely used in Europe.

This study proposes a novel data-driven bedside monitoring approach, where the major goal is to study the impact of adverse events to daily predict the organ condition risk of six systems (i.e. respiratory, coagulation, hepatic, cardiovascular, neurological and renal). The assumption behind our approach is to use only data collected in the last 24 hours of the ICU length of stay. A large database was considered using bedside monitoring data. The input variables included the case mix (i.e. admission type/origin, SAPSII index and the age) and adverse events. The latter were measured as the out of range values of four commonly monitored physiological variables (e.g. heart rate).

The second goal was also to compare two data mining (DM) techniques, namely multinomial logistic regression (MLR) and artificial neural networks (ANNs). The experiments were conducted in the R statistical tool [29] using discrimination and calibration criteria. As argued in [31], it is difficult to compare DM methods in a fair way, with data analysts tending to favor models that they know better. To reduce the bias towards a given model, we adopted the default suggestions of the **nnet** package [15] for the R environment. The only exception is the number of hidden neurons, which was set using a simple grid search procedure. The default settings are more likely to be used by common (non expert) users, thus this seems a reasonable assumption for the comparison.

The results show that the ANNs are the best learning models, outperforming the MLR for both criteria. The average (over all organs) obtained ANN ROC area is 64%, 69% and 74% for the dysfunction, normal and failure conditions, while the respective Brier scores were 0.18, 0.16 and 0.09. In particular, good ANN discrimination results (ROC area higher than 75%) were achieved for three systems (hepatic, neurological and renal). Also, high calibrated models (Brier score below 0.1) were attained for the coagulation, hepatic and renal organs. These results can be explained by the fact that the SOFA score is more reliable and robust when classifying the clinical condition of these organs. For instance, the renal function condition is classified using well defined and objective intervals, rather than respiratory that can be influenced by an inadequate FIO_2 setting.

The risk estimates for the normal and dysfunction conditions provided less accuracies. This may be explained by several factors. Normality is at one the extremes, with the dysfunction being an in-between state. Hence, in principle the normal condition should be easier to predict. However, as shown in Figure 2 there are several outliers (e.g. rare or extreme events) in the data. Since ICU patients are critically ill, the normal function label describes a clinical condition where the severity is not enough to define a failure or dysfunction but does not exclude a disease process. Furthermore, organ failure development is a continuous process where the borders for each stage are necessarily fuzzy and not well known.

Regarding the interpretability issue, the MLR is easier to understand than the neural model. Yet, under the adopted experimental settings, the latter presented the best results and it is possible to extract knowledge from trained ANNs, given in terms of input variable importance or human friendly rules (Section 3.2).

The major outcome of this work is that we show that adverse events, taken from bedside monitored data, have a relevant impact on the degree of organ failure. Although this finding was expected, our main contribution is to quantify such impact (i.e. discrimination, calibration and input relevance), allowing to get knowledge from easy obtainable data. Rather than an empirical subjective analysis (e.g. performed by the individual physician), the obtained results strength the pursuit of a systematic intelligent data-driven approach to monitor ICU patients.

4.1 Related work

In the past, the majority of studies using data mining (DM) methods in ICU environments were focused in mortality assessment [10], while the application of DM to organ failure is rather scarce. Matis et al. [32] used 15 variables (e.g. age, bilirubin, creatinine) to train an ANN in order to predict liver failure after transplantation. The obtained accuracy ranged from 70% (using data prior to the operation) to 88% (5 days after the transplantation). An ANN was also successfully used to access the cardiac failure of 58 patients, using 20 variables (e.g. heart rate, blood pressure) [33]. In previous work [13], ANNs have outperformed decision trees for organ failure prediction, obtaining an overall classification accuracy of 70%. More recently, a kernel logistic regression was used by Pearcea et al. [34] in order to predict acute pancreatitis. The model included 8 variables (e.g. age, respiratory rate, creatinine) and outperformed a daily updated APACHE II prognostic model.

This work is quite distinct from the previous studies, since we use adverse events based on daily bedside monitored data. Moreover, we model the degree of organ failure of six organ systems. This study largely extends our previous work [13] by predicting three conditions (i.e. normal, dysfunction and failure), testing also a logistic model in the experiments and evaluating the results under calibration and discrimination analysis.

Regarding the use of daily SOFA scores by artificial intelligence techniques, most of the literature is also focused on mortality prediction. For instance, Kayaalp et. al [35] adapted bayesian networks under a time series approach, where 23 variables (e.g. urine output, bilirubin, SOFA scores for five organ systems) were used to predict ICU mortality. In previous work [9], we tested the use of ANN and adverse alarms of four biometrics, outperforming the SAPSII logistic model for mortality assessment. Toma et. al [23] followed a distinct dynamic approach, where organ failure scores were used to discover patterns of sequences (called episodes). Several logistic regression models, built for each of the first five days, were tested for mortality prognosis and the best results were attained by the models that included the episodes.

In contrast with the above studies, this work models the degree of organ impairment. Since multiple organ failure is the main cause of ICU mortality, there is a need to identify the degree of ICU patient illness in a continuous form, in order to apply a timely intervention. In fact, this was the rationale behind the SOFA score development [7]. Our study follows a similar and complementary approach, adding a risk estimate (i.e. probability) of the organ condition to bedside alarms. The proposed work could be applied using precise, low cost and real-time variables, by using a real-time computerized data acquisition system from bedside monitors and applying quality procedures (e.g. data validated by the ICU staff) [36]. Moreover, such system could give more updated predictions (e.g. every 6 or 12h).

4.2 Future work

To our knowledge this is the first attempt to related adverse events with organ failure and further exploratory research is needed. For instance, outlier detection techniques [37] could be used to discard rare or extreme cases. This is expected to improve the results, specially for the normal and dysfunction conditions. Moreover, while the adverse events have an impact on organ failure (Section 3.2) there are complex dependencies between the biometrics. Therefore, a temporal analysis, such as presented in [23, 35]. where the evolution of each organ during the patient length of stay is modeled, is a very promising direction. In effect, some of the limitations of this work, namely the manual collection of the data and the lack of temporal sequence analysis, could be answered by testing our approach in a real environment, using real-time data. In effect, we intend to explore all these possibilities in the INTCare pilot project [36], where a friendly decision support system is currently being developed at the ICU of the Hospital Geral de Santo António, Oporto, Portugal.

5 Conclusion

A data-driven analysis was performed on a large ICU database, with an emphasis on the use of daily adverse events, taken from four commonly monitored biometrics. Two data mining methods, artificial neural networks and multinomial logistic regression, were tested to predict the degree of failure regarding six organ systems. The former method provided better discrimination and calibration results, with average ROC curve areas of 74%, 64% and 69% and Brier scores of 0.09, 0.18 and 0.16 for the failure, dysfunction and normal conditions respectively. The obtained results show that adverse events are important intermediate outcomes, reflecting the patient condition and ICU way of work. Hence, this work contributes to an improvement of the process of critical ill patient care, by means of generating more intelligent bedside intensive care alarms.

Acknowledgments

The authors wish to thank FRICE and the BIOMED project BMH4-CT96-0817 for the provision of part of the EURICUS II data, which is integrated in a PhD program, developed at Instituto de Ciências Biomédicas Abel-Salazar from University of Porto and the Departments of Computer Science/Information Systems from the University of Minho. The work of P. Cortez, M.F. Santos and J. Neves is supported by the

FCT project PTDC/EIA/72819/2006. We also would like to thank the anonymous reviewers for their helpful comments.

References

[1] A. Rosenberg. Recent innovations in intensive care unit risk-prediction models. *Current Opinion in Critical Care*, 8:321–330, 2002.

[2] W. Knaus, D. Wagner, E. Draper, J. Zimmerman, M. Bergner, P. Bastos, C. Sirio, D. Murphy, T. Lotring, and A. Damiano. The APACHE III prognostic system. Risk prediction of hospital mortality for critically ill hospitalized adults. *Chest*, 100:1619–1636, 1991.

[3] J. Le Gall, S. Lemeshow, and F. Saulnier. A new simplified acute physiology score (SAPS II) based on a European / North American multicenter study. *JAMA*, 270:2957–2963, 1993.

[4] S. Lemeshow, D. Teres, J. Klar, J. Avrunin, S. Gehlbach, and J. Rapoport. Mortality Probability Models (MPM II) based on an international cohort of intensive care unit patients. *JAMA*, 270:2478–2486, 1993.

[5] J. Le Gall. The use of severity scores in the intensive care unit. *Intensive Care Med*, 31:1618–1623, 2005.

[6] J. Vincent, R. Moreno, J. Takala, S. Willatss, A. Mendonca, H. Bruining, C. Reinhart, P. Suter, and L. Thijs. The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction / failure. *Intensive Care Med*, 22:707–710, 1996.

[7] R. Moreno, J. Vincent, R. Matos, A. Mendonça, F. Cantraine, L. Thijs, J. Takala, C. Sprung, M. Antonelli, H. Buining, and S. Willatts. The use of maximum SOFA score to quantify organ dysfunction/failure in intensive care.

Results of a prospective, multicentre study. *Intensive Care Med*, 25:696–696, 1999.

[8] A. Amaral, F. Andrade, R. Moreno, A. Artigas, F. Cantraine, and J. Vincent. Use of the Sequential Organ Failure Assessment score as a severity score. *Intensive Care Med*, 31:243–249, 2005.

[9] Á. Silva, P. Cortez, M. F. Santos, L. Gomes, and J. Neves. Mortality assessment in intensive care units via adverse events using artificial neural networks. *Artif Intell Med*, 36:223–234, 2006.

[10] L. Ohno-Machado, F. Resnic, and M. Matheny. Prognosis in Critical Care. *Annual Rev Biomed Eng*, 8:567–599, 2006.

[11] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, MA, 2001.

[12] S. Haykin. *Neural Networks - A Compreensive Foundation*. Prentice-Hall, New Jersey, 2nd edition, 1999.

[13] Á. Silva, P. Cortez, M. Santos, L. Gomes, and J. Neves. Multiple Organ Failure Diagnosis Using Adverse Events and Neural Networks. In I. Seruca, J. Cordeiro, S. Hammoudi, and J. Filipe, editors, *Enterprise Information Systems VI*, The Netherlands, 2006. Springer.

[14] V. Fidler, R. Nap, and R. Miranda. The effect of a managerial-based intervention on the occurrence of out-of-range-measurements and mortality in Intensive Care Units. *Journal of Critical Care*, 19(3):130–134, 2004.

[15] W. Venables and B. Ripley. *Modern Applied Statistics with S*. Springer, 4th edition, 2003.

[16] R. Setiono. Techniques for Extracting Classification and Regression Rules from Artificial Neural Networks. In D. Fogel and C. Robinson, editors, *Computational Intelligence: The Experts Speak*, pages 99–114. Piscataway, NY, USA, IEEE, 2003.

[17] L. Breiman, J. Friedman, R. Ohlsen, and C. Stone. *Classification and Regression Trees*. Wadsworth, Monterey, CA, 1984.

[18] R. Kewley, M. Embrechts, and C. Breneman. Data Strip Mining for the Virtual Design of Pharmaceuticals with Neural Networks. *IEEE Trans Neural Networks*, 11(3):668–679, May 2000.

[19] P. Cortez, M. Portelinha, S. Rodrigues, V. Cadavez, and A. Teixeira. Lamb Meat Quality Assessment by Support Vector Machines. *Neural Processing Letters*, 2006.

[20] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.

[21] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27:861–874, 2006.

[22] F. Provost and P. Domingos. Tree Induction for Probability-Based Ranking. *Machine Learning*, 52(3):199–215, 2003.

[23] T. Toma, A. Abu-Hanna, and R. Bosman. Discovery and inclusion of SOFA score episodes in mortality prediction. *Journal of Biomedical Informatics*, 40(6):649–660, 2007.

[24] J. Bi and K. Bennett. Regression Error Characteristic curves. In T. Fawcett and N. Mishra, editors, *Proceedings of 20th Int. Conf. on Machine Learning (ICML)*, Washington DC, USA, AAAI Press, 2003.

[25] R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Volume 2, Montreal, Quebec, Canada, Morgan Kaufmann, August 1995.

[26] K. Molodianovitch, D. Faraggi, and B. Reiser. Comparing the Areas Under Two Correlated ROC Curves: Parametric and Non-Parametric Approaches. *Biometrical Journal*, 48(5):745–757, 2006.

[27] E. DeLong, D. DeLong, and D. Clarke-Pearson. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Non-parametric Approach. *Biometrics*, 44(3):837–845, 1988.

[28] P. Cortez. RMiner: Data Mining with Neural Networks and Support Vector Machines using R. In R. Rajesh (Ed.), *Introduction to Advanced Scientific Softwares and Toolboxes*, *in press.*

[29] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-00-3, http://www.R-project.org, (Accessed 26 March 2008).

[30] M. Moller. A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, 6(4):525–533, 1993.

[31] D. Hand. Classifier technology and the illusion of progress. *Statistical Science*, 21(1):1–15, 2006.

[32] S. Matis, H. Doyle, I. Marino, R. Mural, and E. Uberbacher. Use of neural networks for prediction of graft failure following liver transplantation. In *Proceedings of the 8th IEEE Symposium on Computer-Based Medical Systems*, pages 133–140, Washington, DC, USA, 1995. IEEE.

[33] M. Gils, H. Jansen, K. Nieminen, R. Summers, and P. Weller. Using artificial neural networks for classifying ICU patient states. *Engineering in Medicine and Biology Magazine*, 16:41–47, 1997.

[34] C. Pearcea, S. Gunn, A. Ahmeda, and C. Johnson. Machine Learning Can Improve Prediction of Severity in Acute Pancreatitis Using Admission Values of APACHE II Score and C-Reactive Protein. *Pancreatology*, 6:123–131, 2006.

[35] M. Kayaalp, G. Cooper, and G. Clermont. Predicting ICU Mortality: A Comparison of Stationary and Nonstationary Temporal Models. In *Proceedings of AMIA Symposium*, pages 418–422, Los Angeles CA, USA, AMIA, 2000.

[36] P. Gago, M.F. Santos, Á. Silva, P. Cortez, J. Neves, and L. Gomes. INTCare: A Knowledge Discovery based Intelligent Decision Support System for Intensive Care Medicine. *Journal of Decision Systems*, 14(3):241–259, 2005.

[37] V. Hodge and J. Austin. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004.

Table 1: The protocol for the out of range physiologic measurements

	BP	**SpO_2**	**HR**	**UR**
Normal Range	$90 - 180$mmHg	$\geq 90\%$	$60 - 120$bpm	≥ 30ml/h
Event[a]	≥ 10min.	≥ 10min.	≥ 10min.	≥ 1h
Event[b]	≥ 10min. in 30min.	≥ 10min. in 30min.	≥ 10min. in 30min.	–
Critical Event[a]	≥ 1h	≥ 1h	≥ 1h	≥ 2h
Critical Event[b]	≥ 1h in 2h	≥ 1h in 2h	≥ 1h in 2h	–
Critical Event[c]	< 60mmHg	$< 80\%$	< 30bpm $\vee > 180$bpm	≤ 10ml/h

BP - blood pressure, HR - heart rate, SpO_2 - pulse oximeter oxygen saturation, UR - urine output.

a Defined when continuously out of range.

b Defined when intermittently out of range.

c Defined anytime.

Table 2: The intensive care variables

Attribute	**Description**	**Min**	**Max**	**Mean**[a]
admtype	admission type	Categorical[b]		
admfrom	admission origin	Categorical[c]		
SAPS II	SAPS II score	0	118	40.9±16.4
age	age of the patient	18	100	62.5±18.2
NBP	daily number of blood pressure events	0	24	0.8±1.9
NHR	daily number of heart rate events	0	24	0.6±2.3
NSpO$_2$	daily number of oxygen events	0	24	0.4±1.8
NUR	daily number of urine events	0	24	1.0±3.0
NCRBP	daily number of critical blood pressure events	0	10	0.3±0.7
NCRHR	daily number of critical heart rate events	0	10	0.2±0.6
NCRSpO$_2$	daily number of critical oxygen events	0	6	0.1±0.4
NCRUR	daily number of critical urine events	0	7	0.4±0.8
TCRBP	time of critical blood pressure events (% of 24h)	0	24.7	0.8±2.7
TCRHR	time of critical heart rate events (% of 24h)	0	24.7	1.0±3.4
TCRSpO$_2$	time of critical oxygen events (% of 24h)	0	24.7	0.4±2.1
TCRUR	time of critical urine events (% of 24h)	0	24.7	1.6±4.5

a mean and sample standard deviation.

b 1 - unscheduled surgery, 2 - scheduled surgery, 3 - medical.

c 1 - operating theatre, 2 - recovery room, 3 - emergency room, 4 - general ward, 5 - other ICU, 6 - other hospital, 7 - other sources.

Table 3: The SOFA variables and scoring rules (adapted from [7])

Organ/			**SOFA Score**		
Variable	0	1	2	3	4
respiratory					
PaO_2/FIO_2 (mmHg)	>400	≤400	≤ 300	$\leq 200^a$	$\leq 100^a$
coagulation					
platelets$\times 10^3/mm^3$	>150	≤150	≤ 100	≤ 50	≤ 20
hepatic					
bilirubin (μmol/l)	>20	<32	< 101	< 204	> 204
cardiovascular					
hypotension^b	None	MAP< 70 mmHg	dop.≤5 or dobutamine (any dose)	dop.<5 or epi.≤0.1 or norepi.≤0.1	dop.>15 or epi. > 0.1 or norepi.>0.1
neurological					
Glasgow coma score	15	13-14	10-12	6-9	<6
renal					
creatinine (μmol/l)	<110	≥110	≥ 171	≥ 300	≥ 440
or urine output				<500mL/day	<200ml/day

PaO_2 - arterial oxygen tension, FIO_2 - fractional inspired oxygen.

MAP - mean arterial pressure, dop. - dopamine, epi. - epinephrine,

norepi. - norepinephrine.

a – with respiratory support.

b – agents administered for at least 1 hour (doses in μg/kg per min).

Table 4: The discrimination power (mean AUC value of the 20 runs, in percentage) for each organ, condition and method (values of AUC>70% are in bold)

	Normal		Dysfunction		Failure		Global	
Organ	MLR	ANN	MLR	ANN	MLR	ANN	MLR	ANN
respiratory	67.2	69.5	59.2	61.0	65.6	68.9	63.6	66.0
coagulation	63.6	65.5	60.1	62.0	**72.6**	**73.9**	63.3	65.1
hepatic	64.7	66.7	62.5	64.2	**72.6**	**76.0**	64.6	66.6
cardiovascular	67.9	**71.2**	63.8	65.6	67.3	**71.0**	67.1	**70.2**
neurological	**70.0**	**72.1**	58.8	61.2	**74.7**	**76.7**	68.8	**70.9**
renal	69.4	**70.7**	66.0	66.8	**73.5**	**76.1**	69.1	**70.4**
Average	67.1	69.3	61.7	63.5	**71.0**	**73.8**	66.1	68.2

Table 5: The calibration values (mean Brier score of the 20 runs) for each organ, condition and method (values in bold denote statistical significance when compared with MLR)

	Normal		Dysfunction		Failure		Global	
Organ	MLR	ANN	MLR	ANN	MLR	ANN	MLR	ANN
respiratory	0.213	**0.204**	0.233	**0.230**	0.171	**0.166**	0.211	**0.205**
coagulation	0.173	**0.171**	0.155	**0.154**	0.038	0.038	0.134	**0.133**
hepatic	0.132	**0.130**	0.116	0.116	0.026	**0.025**	0.101	**0.100**
cardiovascular	0.205	**0.197**	0.132	**0.130**	0.138	**0.133**	0.160	**0.155**
neurological	0.208	**0.202**	0.153	**0.151**	0.136	**0.132**	0.169	**0.165**
renal	0.182	**0.179**	0.155	**0.155**	0.065	**0.063**	0.144	**0.142**
Average	0.185	**0.181**	0.157	**0.156**	0.096	**0.093**	0.153	**0.150**

Table 6: The relative importance of the input variables for the multinomial logistic regression (R_a values, in percentage).

Organ	admtype	admfrom	SAPS II	age	BP*	HR*	SpO_2*	UR*
respiratory	17.7	4.6	14.1	6.3	12.5	6.8	34.4	3.6
coagulation	16.5	9.3	12.5	6.1	15.0	9.1	20.9	10.6
hepatic	8.0	11.6	5.9	4.7	8.2	37.4	10.1	14.1
cardiovascular	2.3	16.0	22.6	6.6	11.2	19.9	8.3	13.1
neurological	4.1	14.9	22.7	4.8	10.5	20.5	19.0	3.5
renal	5.9	4.3	16.6	10.1	20.7	17.0	11.9	13.5
Average	9.1	10.1	15.7	6.5	13.0	18.5	17.4	9.7

* – All attributes related to the variable where summed (number of events, critical events and the time).

Table 7: The relative importance of the input variables for the artificial neural networks (R_a values, in percentage).

Organ	admtype	admfrom	SAPS II	age	BP*	HR*	SpO_2*	UR*
respiratory	16.8	7.8	15.1	10.0	19.9	8.1	17.1	5.2
coagulation	30.9	10.8	12.7	7.0	7.5	2.6	18.1	10.4
hepatic	23.1	7.8	12.1	10.8	9.1	5.1	17.0	15.0
cardiovascular	14.1	17.3	16.5	12.8	9.8	9.6	13.4	6.5
neurological	31.2	10.2	15.6	7.5	17.3	3.5	10.4	4.3
renal	2.3	13.6	26.6	9.9	5.1	6.4	19.8	16.3
Average	19.7	11.3	16.4	9.7	11.4	5.9	16.0	9.6

* – All attributes related to the variable where summed (number of events, critical events and the time).

Table 8: The multinomial logistic coefficients for the renal system.

Condition	$\beta i,j$ coefficients
failure	$-0.32 - 0.50 admtype_2 + 0.10 admtype_3 + 0.14 admfrom_2 + 0.11 admfrom_3$ $+0.13 admfrom_4 + 0.51 admfrom_5 + 0.03 admfrom_6 - 0.04 admfrom_7$ $+0.01 SAPSII - 0.02 age - 0.05 NBP - 0.05 NCRBP - 0.01 NHR$ $-0.17 NCRHR - 0.03 NSpO_2 + 0.09 NCRSpO_2 - 0.03 NUR$ $+0.03 TCRBP - 0.03 TCRHR - 0.06 TCRSpO_2 + 0.12 TCRUR$
normal	$3.56 - 0.20 admtype_2 - 0.05 admtype_3 - 0.11 admfrom_2 + 0.15 admfrom_3$ $+0.15 admfrom_4 - 0.05 admfrom_5 + 0.18 admfrom_6 + 0.55 admfrom_7$ $-0.03 SAPSII - 0.02 age - 0.04 NBP - 0.13 NCRBP - 0.01 NHR$ $-0.12 NCRHR + 0.04 NSpO_2 - 0.15 NCRSpO_2 + 0.06 NUR$ $+0.01 TCRBP - 0.02 TCRHR - 0.01 TCRSpO_2 - 0.07 TCRUR$

Binary variables are denoted by V_i, denoting the i-th categorical value of variable V.

List of figure captions:

Figure 1. The organ condition prevalence during the ICU length of stay (x-axis denotes the daily SOFA value and the y-axis the frequency of the x value within the whole dataset).

Figure 2. Boxplots of the time of critical events for each renal condition. Each box is delimited by first (bottom) and third (top) quartiles. Mean values are represented by black diamonds and outliers by open circles. The latter were defined if outside $1.5\times$ the interquartile range of the box.

Figure 3. Example of a multinomial logistic regression (left) and artificil neural network with 2 hidden nodes (right).

Figure 4. The receiver operating characteristic curves for each organ and condition (artificial neural network – solid line, multinomial logistic regression – dashed, random – gray line).

Figure 5. The regression error curves for each organ and condition (artificial neural network – solid line, multinomial logistic regression – dashed).

Figure 6. The extracted rules given in terms of a decision tree for the renal system.

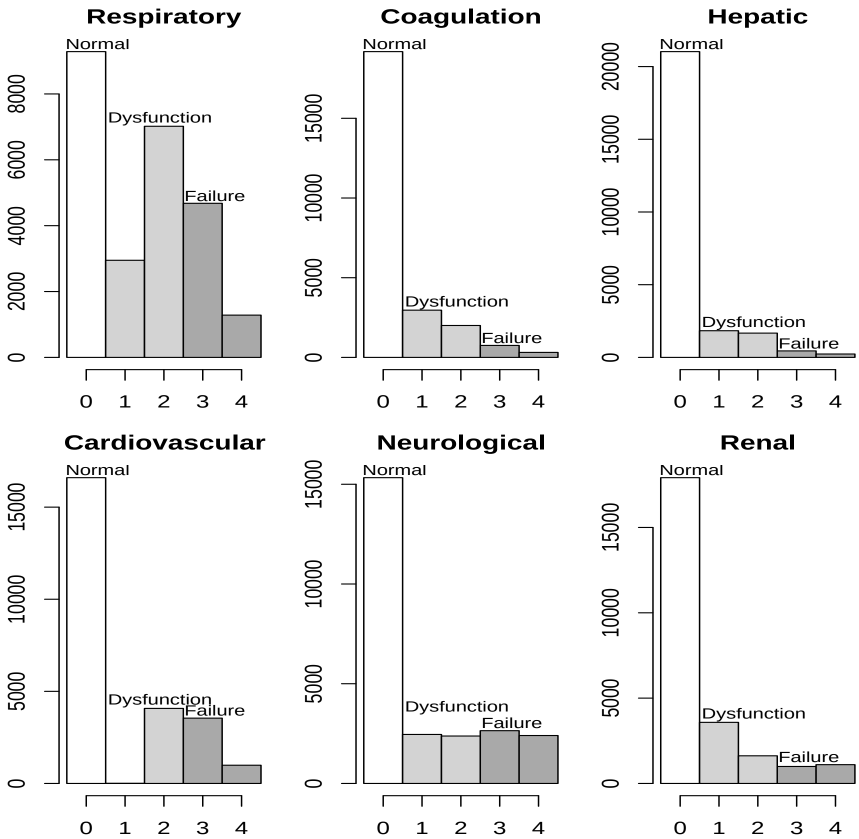

Respiratory
Normal
Dysfunction
Failure
0
2000
4000
6000
8000
0 1 2 3 4
Coagulation
Normal
Dysfunction
Failure
0
5000
10000
15000
0 1 2 3 4
Hepatic
Normal
Dysfunction
Failure
0
5000
10000
15000
20000
0 1 2 3 4
Cardiovascular
Normal
Dysfunction
Failure
0
5000
10000
15000
0 1 2 3 4
Neurological
Normal
Dysfunction
Failure
0
5000
10000
15000
0 1 2 3 4
Renal
Normal
Dysfunction
Failure
0
5000
10000
15000
0 1 2 3 4

Figure 1:

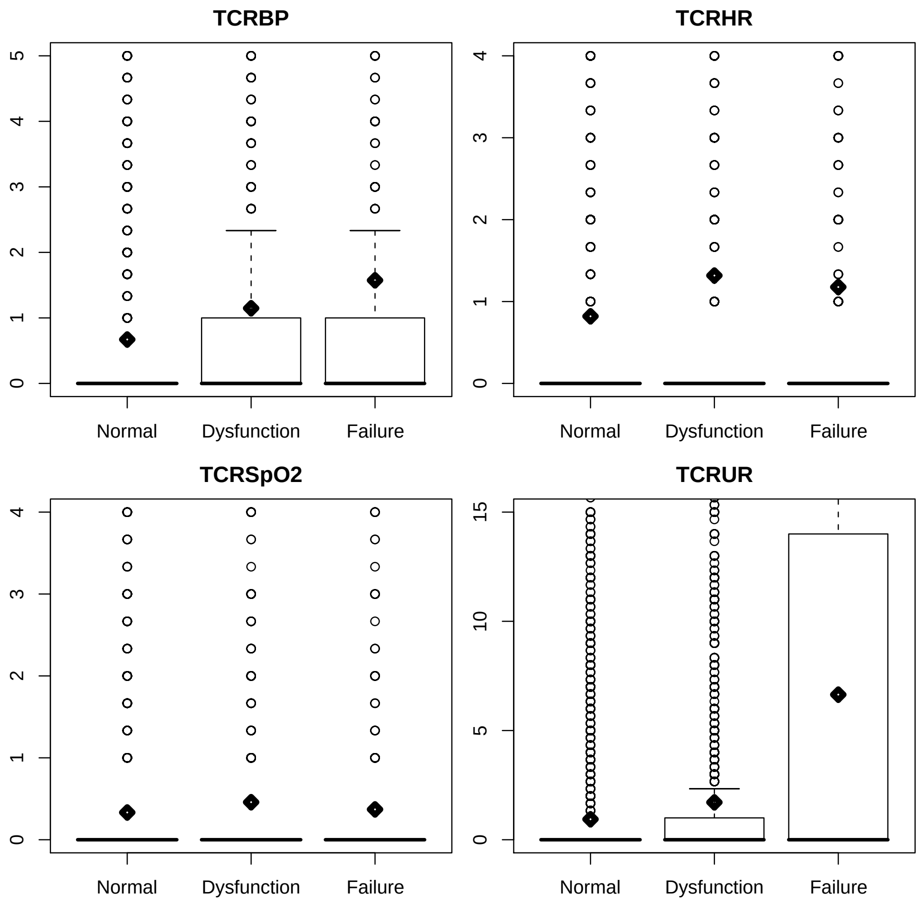

TCRBP
TCRHR
TCRSpO2
TCRUR
Normal
Dysfunction
Failure
0
1
2
3
4
5
5
10
15

Figure 2:

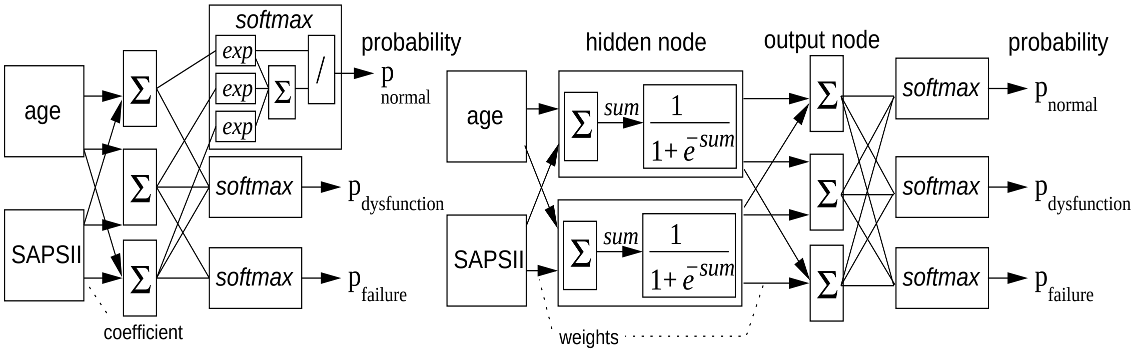

softmax
exp
exp
exp
Σ
/
probability
p normal
age
SAPSII
Σ
Σ
Σ
softmax
p dysfunction
softmax
p failure
coefficient
hidden node
output node
probability
age
SAPSII
Σ
sum
1
1+ e−sum
Σ
sum
1
1+ e−sum
Σ
Σ
Σ
softmax
p normal
softmax
p dysfunction
softmax
p failure
weights

Figure 3:

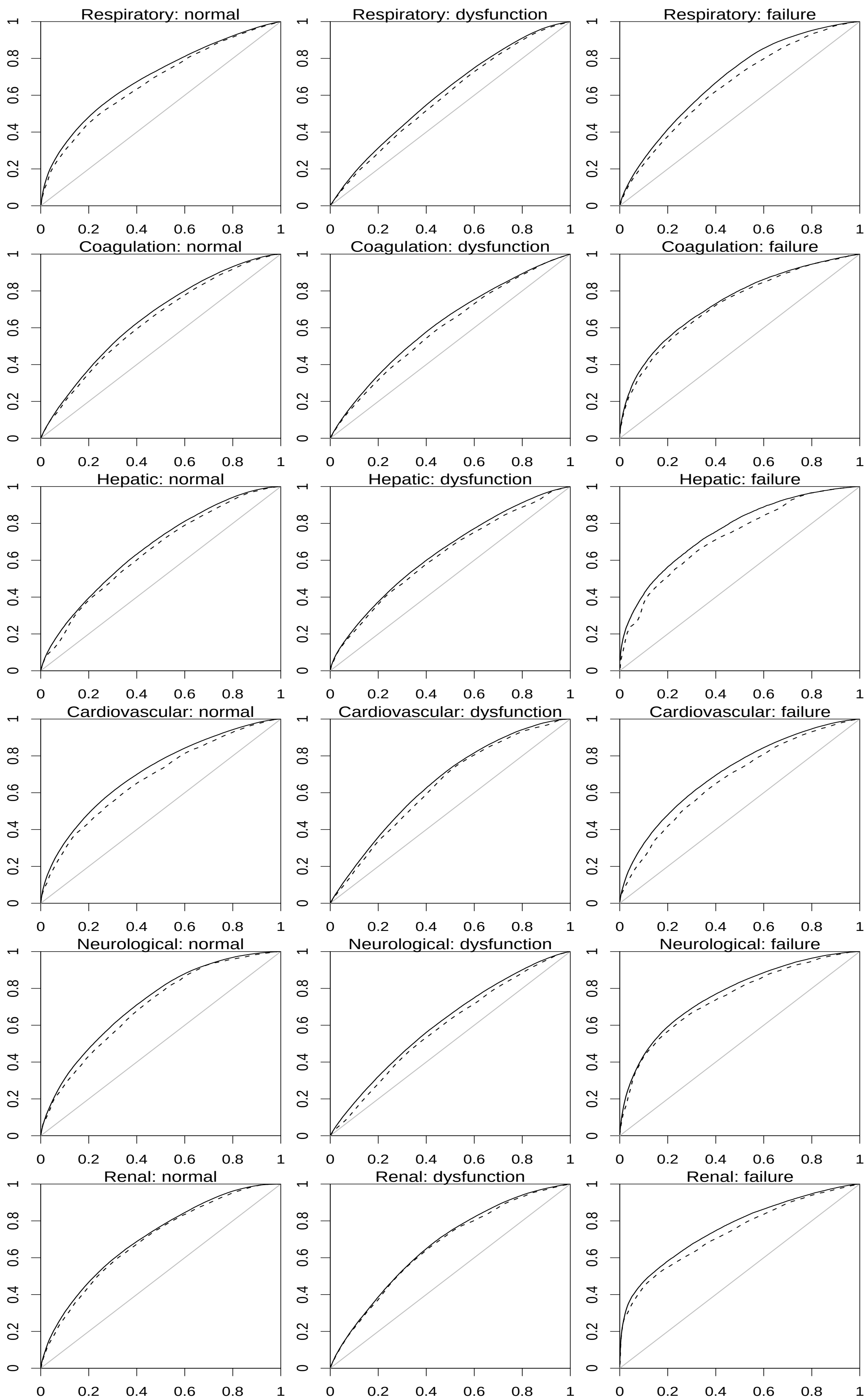
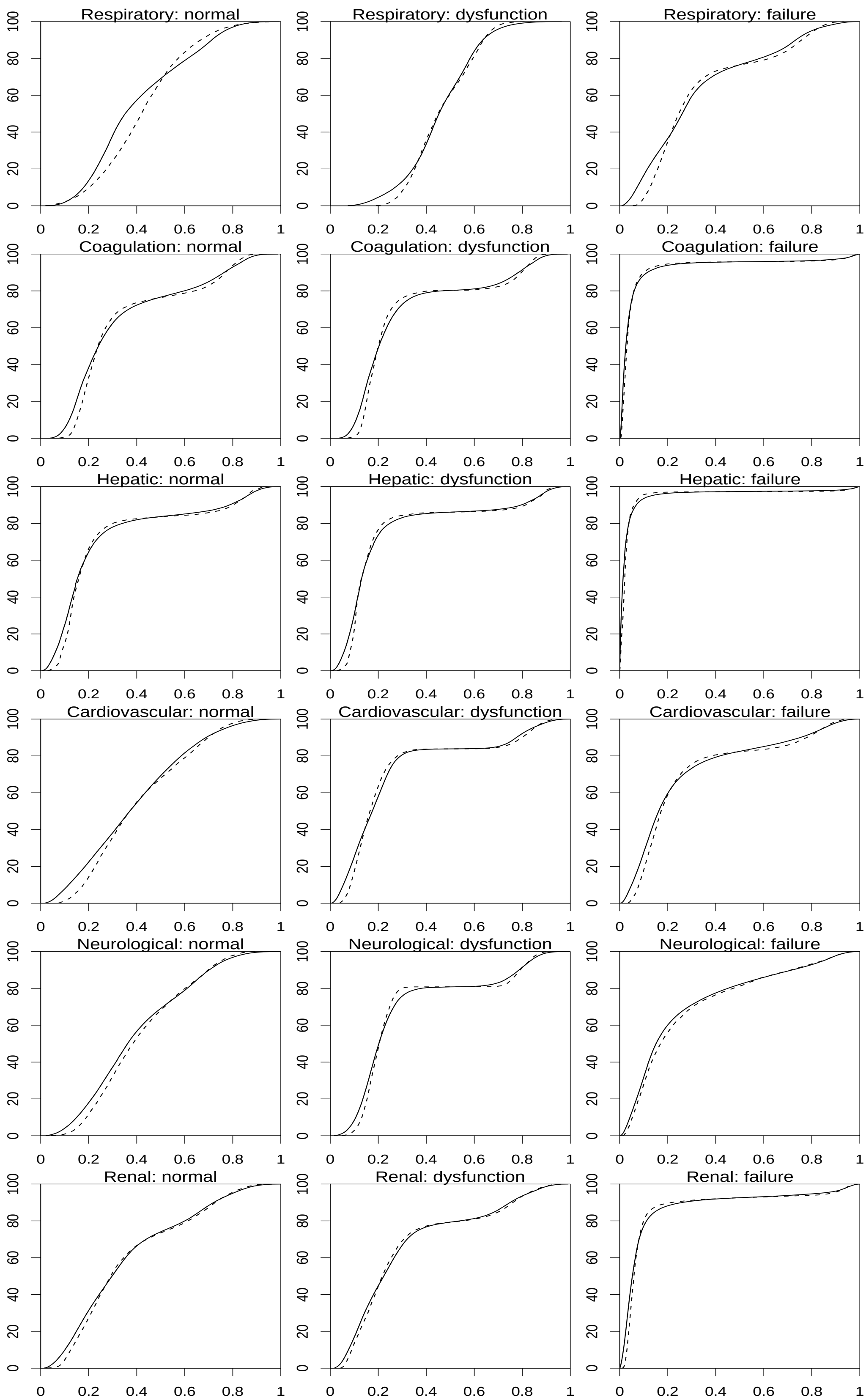

Respiratory: normal
Respiratory: dysfunction
Respiratory: failure
Coagulation: normal
Coagulation: dysfunction
Coagulation: failure
Hepatic: normal
Hepatic: dysfunction
Hepatic: failure
Cardiovascular: normal
Cardiovascular: dysfunction
Cardiovascular: failure
Neurological: normal
Neurological: dysfunction
Neurological: failure
Renal: normal
Renal: dysfunction
Renal: failure
0
0.2
0.4
0.6
0.8
1

Figure 4:



Respiratory: normal
Respiratory: dysfunction
Respiratory: failure
Coagulation: normal
Coagulation: dysfunction
Coagulation: failure
Hepatic: normal
Hepatic: dysfunction
Hepatic: failure
Cardiovascular: normal
Cardiovascular: dysfunction
Cardiovascular: failure
Neurological: normal
Neurological: dysfunction
Neurological: failure
Renal: normal
Renal: dysfunction
Renal: failure

Figure 5:

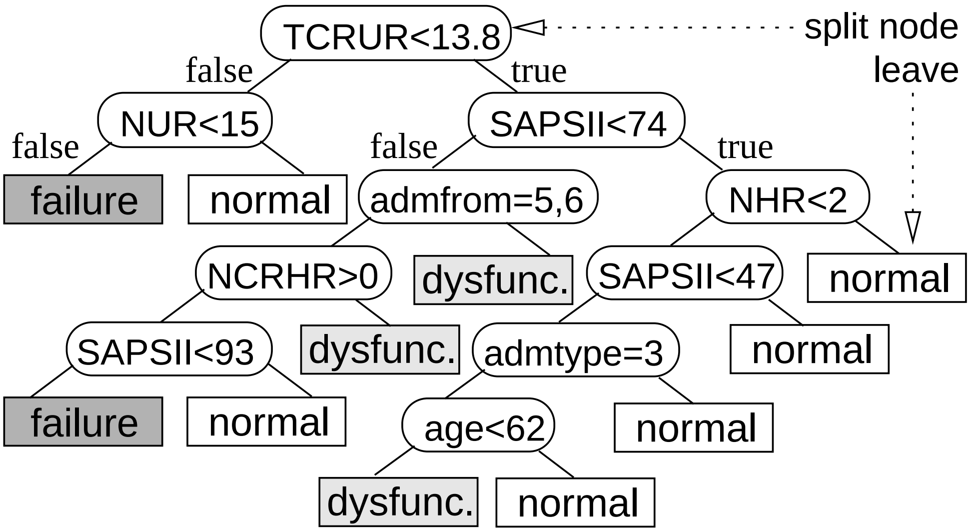

TCRUR<13.8
split node
leave
false
true
NUR<15
SAPSII<74
false
false
true
failure
normal
admfrom=5,6
NHR<2
NCRHR>0
dysfunc.
SAPSII<47
normal
SAPSII<93
dysfunc.
admtype=3
normal
failure
normal
age<62
normal
dysfunc.
normal

Figure 6: