Jurnal
Teknologi

Full Paper

DATA MINING TO PREDICT THE USE OF VASOPRESSORS IN INTENSIVE MEDICINE PATIENTS

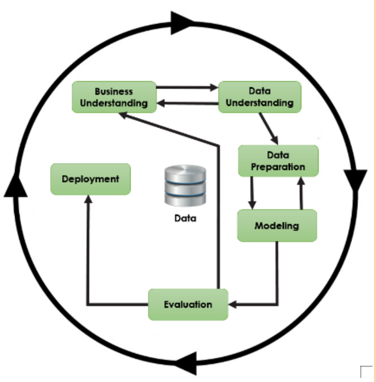
André Braga[a], Filipe Portela[a,b*], Manuel Filipe Santos[a], António Abelha[a], José Machado[a], Álvaro Silva[c], Fernando Rua[c]

[a]Algoritmi Research Centre, University of Minho, Portugal
[b]ESEIG, Porto Polytechnic, Vila do Conde, Portugal
[c]Intensive Care Unit, Centro Hospitalar do Porto, Portugal

Article history
Received
1 June 2015
Received in revised form
15 September 2015
Accepted
26 December 2015

*Corresponding author
cfp@dsi.uminho.pt

Graphical abstract



Abstract

The role that decision making process plays in Intensive Medicine is very critical essential due to the bad health condition of the patients that go to Intensive Care Units (ICU) and the need of a quick and accurate decisions. Therefore each decision is crucial, because it can help saving endangered lives. The decision should be always taken in the patient best interest after analyzing all the data available. In the eyes of the intensivists, the ever growing amount of available data concerning the patients, makes it each time more difficult for them to make a decision based on so many information. It is based on this ideal of improving the decision making process, that this work arises and Data Mining models were induced to predict if a patient will need to take a vasopressor, more specifically: Dopamine, Adrenaline or Noradrenaline. This work used real data provided by an Intensive Care Unit and collected in real-time. The data mining model were induced using data from vital sign monitors, laboratory analysis and information about the patient's Electronic Health Record. This study was based in clinical evidences and provided very useful results with a sensitivity around 90%. These models will reduce the need of vasopressor drugs by helping intensivists to act and take accurate decision before the vasopressor be need by the patient. It will improve the patient condition because when the time comes the predicted necessity of the vasopressor will cease to exist due to the early care provided by the intensivist. The decisions can be for example change the therapeutic plan. Overall, the decision making process becomes more reliable and effective and the quality of care given to patients is better.

Keywords: Intensive medicine, intensive care units, patient-centered, data mining; intcare, vasopressor, decision making

1.0 INTRODUCTION

Nowadays, it is possible to observe how Information Technologies (IT) intervenes in the most diverse areas in order to provide knowledge not available till the moment. The field of Intensive Medicine (IM) is no exception, and it has been put through several changes due to the implementation of IT, which have one main goal: improve the quality of care given to patients [1]. There are many ways in which IT can be applied, being one of them the use of Data Mining (DM) techniques. The implementation of technology in Intensive Care Units (ICU) increased the amount of data generated, and so DM seeks to transform the large amounts of data into new knowledge that will support the decision making process in ICU [2].

This study aims to create new knowledge that will help the decision making process of intensivists working in ICU. Through the use of DM techniques several data mining models were induced. The main goal is to be able to early predict whether a patient will need to take a vasopressor. If it holds true the system will alert the intensivists for this future necessity, allowing them to act before and make that necessity cease to exist. The data used to make the predictions holds records relative to patients' lab analysis, vital signs and Electronic Health Record (EHR) and is provided by the ICU of the Centro Hospitalar do Porto (CHP). After applying an Extract, Transform and Loading (ETL) process to the data, were then developed the scenarios. The best scenario hold a result of 90.72% for the sensitivity metric, which according to the clinical context is a very interesting and satisfactory result.

Besides the introduction, this article is composed by other four sections. Background is the second section and holds background knowledge about the development of this study, such as Intensive Medicine and Intensive Care Units, Vasopressors, the INTCare system and Data Mining. In section three Study Description are enunciated the methods and tools used in the development of the study along with the phases of the CRISP-DM methodology. The fourth section is where are discussed the results of the study. Lastly, in the fifth chapter are drawn the conclusions regarding the study and it's enunciated on what future work will focus on.

2.0 BACKGROUND

2.1 Intensive Medicine and Intensive Care Units

Intensive Medicine (IM) can be seen as a multidisciplinary field of medical sciences which focus on addressing three main aspects: prevention, diagnosis and treatment of patients with reversible pathophysiological conditions. It reveals to be a threat to patients' lives by making their vital functions cease to work properly [3]. The process of reversing a patient's condition is done in qualified facilities named Intensive Care Units (ICU). Here specifically trained physicians named intensivists look after the patients and through the use of drugs and with the support of life-support devices to continuously monitoring the patients' vital signs. They make sure that the patient regain their previous health condition [4].

2.2 Vasopressors

A vasopressor is a type of drug that has as main function the increase of blood pressure. This study will focus on three sympathomimetic drugs, i.e., drugs that mimic the effects of neurotransmitter substances of the sympathetic nervous system. These drugs are Adrenaline, Noradrenaline and Dopamine.

Members of the catecholamine family of neurotransmitters, adrenaline, noradrenaline and dopamine act as agonists in α_1 and α_2 adrenergic receptors, resulting in vasoconstriction and therefore increase in blood pressure [5]. Adrenaline is usually used to treat cardiac arrest, asthma and anaphylaxis, while dopamine and noradrenaline can be used to treat hypotension, cardiac arrest and septic shock [6-8]. The used variables for the Data Mining models were selected based on the possible causes of these illnesses that end up requiring vasopressors to be treated. For example Adrenaline can be used to treat a cardiac arrest, which has as possible causes the lack of oxygen, abnormal pH in the body, excess or scarcity of potassium, low body temperature, low or high glucose, among others [9]. Dopamine is often used to treat acute hypotension which is a result of low blood pressure. Noradrenaline can be used to treat sepsis which has as possible causes the increase in heart rate, breathing rate, leucocytes and low oxygen [10, 11]. Even though the blood pressure increase is the most important effect for the context of this study, these vasopressors can also have many other different effects, which also depend on the dosage level.

2.3 INTCare

INTCare is a research project that sought to evolve the process through which data was acquired by the Centro Hospitalar do Porto's (CHP) information system in order to make them available to be used by the decision process. The ICU information system was changed from a manually and on paper data registration to an automatic and electronic process of data acquisition in real-time. With these changes the information system gained the ability to be more proactive, without giving only a reactive response [12].

Implemented in the Intensive Care Unit (ICU) of CHP, the INTCare system is a Pervasive Intelligent Decision Support System (PIDSS) [13, 14] [15] seeking to support the decision making process of the intensivists. It does so by providing new knowledge in real-time about clinic events, such as predicting patient's outcome, organ failure [16], readmission [17], discharge [18], among others [19-21]. The system is based on intelligent agents interacting between them and belong to four subsystems: Data Acquisition, Knowledge Management, Inference and Interface [22]. The tasks performed by the agents do not require human intervention , and so they automate the collection, processing and transformation of data [23], inducing in real-time the data mining models [19, 24]. With this real-time collection of information [25] the system is capable of detect and alert when a patient's vital function leave their normal range [13].

These characteristics have been defined in regard to the environment [15], the system and the information that is available, in order to make sure that the results are appropriated [19]. This facilitates the decision making process in the ICU, making them more effective and efficient environments, and more important, more capable of treating patients. INTCare project is in a second stage where the therapeutic plans and the risks associated to them are one of the concerns. In this sense early predicting the use of Vasopressors is one of the goals.

2.4 Data Mining

In general, Data Mining (DM) consists in converting data into useful information. More specifically, DM is a process of looking for patterns and relationships within large amounts of data, with the purpose to predict or describe. The description aspect is focused on looking for patterns, presenting them in an understandable way thereafter, while the prediction aspect seeks to identify variables in the database that may be useful to predict future events [26].

There are six commonly known type of tasks performed when doing data mining: anomaly detection, association rule learning, clustering, classification, regression and summarization [27]. This study focused on a classification oriented approach.

A classification process seeks to find common properties within a set of data in a database, and classifies them into different classes [28]. For example, this study sought to classify patients that took, at least, one of the named vasopressors and patients who did not take any of the vasopressors.

3.0 STUDY DESCRIPTION

3.1 Method and Tools

Cross Industry Standard Process for Data Mining (CRISP-DM) was the methodology used in this study. CRISP-DM is a standard approach to solve data mining problems and it is divided in 6 phases order by: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment [29]. Even though it has an order to be followed it is possible and advised to iterate on previous phases, in order to improve the final results.

For the development of this study an Integrated Development Environment was used: Oracle SQL Developer for data analysis, understanding and preparation, along with its Oracle Data Miner (ODM) extension to build the prediction models. In modeling, all the techniques provided by ODM were used: Support Vector Machines (SVM), Decision Trees (DT), Naive Bayes (NB) and Generalized Linear Model (GLM).

3.2 Business and Data Understanding

The main purpose of this study is to create data mining models capable of predicting whether a patient will need vasopressor or not, based on patient's clinical information. These models' intent is to improve the quality of care provided to patients, as well as help intensivists make decisions with more precision.The data used in the study was provided by Centro Hospitalar do Porto's (CHP) database, containing data concerning patient's laboratory analysis, vital signs and Electronic Health Record (EHR). The timeframe of the data used is encompassed between January 6th, 2015 and May 18th, 2015 which corresponds to 1259 rows of information about 56 distinct patients. Seventeen variables were considered: SPO2, ECG_HR, ART_SYS, TEMP_T1, PH, Erythrocytes, Potassium, Glucose, Leucocytes, Lactate, PCO2, Hemoglobin, Age, Sex, Provenience, Type_Hospitalization and Hospitalization_Surgery. VSPGeral is the models target. Below in Table 1, it is an overview about the used variables.

Table 1 Variables Overview

Variable	Type	Source	Max	Avg	Min	Standard Deviation	Variance
SPO2	Number	Vital Signs	100	44.59	95.82	6.22	38.64
ECG_HR	Number	Vital Signs	187	83.19	39	17.93	321.50
ART_SYS	Number	Vital Signs	238.42	129.74	238.42	28.75	826.71
TEMP_T1	Number	Vital Signs	39.63	36.48	34	1.03	1.06
PH	Number	Lab Analysis	7.80	7.43	1.60	0.08	0.008
Erythrocytes	Number	Lab Analysis	5.16	3.24	1.17	0.54	0.29
Potassium	Number	Lab Analysis	7.60	3.78	1.60	0.76	0.58
Glucose	Number	Lab Analysis	389	143.69	14	48.58	2360.66
Variable	**Type**	**Source**	**Max**	**Avg**	**Min**	**Standard Deviation**	**Variance**
Leucocytes	Number	Lab Analysis	53.18	12.11	1.24	7.92	62.76
Lactate	Number	Lab Analysis	16	2.35	0.30	2.33	5.45
PCO2	Number	Lab Analysis	172	41.78	13.70	12.27	150.61
Hemoglobin	Number	Lab Analysis	20.60	9.81	2	2.03	4.12
Age	Number	EHR	90	62.78	21	15.50	240.29
Sex	String	EHR	-	-	-	-	-
Provenience	String	EHR	-	-	-	-	-
Type_Hospitalization	String	EHR	-	-	-	-	-
Hospitalization_Surgery	String	EHR	-	-	-	-	-
VSPGeral	String	-	-	-	-	-	-

3.3 Data Preparation

In the data preparation phase an Extract, Transform and Loading (ETL) process was conducted. This process took the existing data through various iterations until it had the desired quality state. Most variables had corrections due to wrong or mistyped information. Wrong or missing information was deleted and mistyped information was merely corrected. Other variables, such as Age and VSPGeral were derived from other variables. Age was derived from date of birth and VSPGeral is the result of the combination of three variables: Dopamine, Adrenaline and Noradrenaline. Whenever at least one of these three had been taken. VSPGeral's value was registered as 1, otherwise as 0.

Having finished the ETL process the final result of the distribution of the target variable (VSPGeral) can be seen below in Figure 1. We can see that 37,81% of total data represents records in which, at least one of the three vasopressors was taken, while 62,19% represent the records which have no registry of a vasopressor being taken.

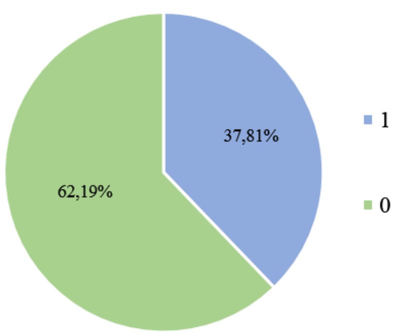


Figure 1 VSPGeral Value Distribution

3.4 Modeling

In this phase, the focus is to create models that through the use of data mining techniques will return results concerning the objective of the study. In order to better understand how the modeling was done, four groups of variables were created. These groups were used to predict the target variable.
Below it is possible to see which groups were created, as well as the variables that make up each of the groups:

Target Variable = VSPGeral
Vital Signs (VS) = SPO2, ECG_HR, ART_SYS, TEMP_T1
Lab Analysis (LA) = PH, Erythrocytes, Potassium, Glucose, Leucocytes, Lactate, PCO2, Hemoglobin
Patient Admission (PA) = Provenience, Type_Hospitalization, Hospitalization_Surgery
Case Mix (CM) = Age, Sex

Based on the formed groups of variables, 6 scenarios were modeled, in which all have used 60% of data for training, while the remaining 40% were used for test. The difference here is that the training block focuses on producing the data mining model and the testing block looks to measure the ability of generalization of the model being produced. In Table 2, are represented the modeled scenarios and which groups of variables each of the scenarios hold.

Table 2 Scenarios and their Group of Variables

Scenarios	Groups
S1	VS + LA + PA + CM
S2	VS + LA + PA
S3	VS + LA + CM
S4	LA + CM
S5	VS + CM
S6	PA + CM

Combining all scenarios, a total of 24 models were developed (6 Scenarios * 4 Techniques). These can be represented by the following expression:

$$M_n = A_x + S_i + T_z$$

In the expression M_n is the model with a classification approach (A_x), a scenario (S_i) and a technique (T_z):

$A_x = \{Classification\}$
$S_i = \{Scenario\ 1, \dots, Scenario\ 6\}$
$T_z = \{SVM, NB, GLM, DT\}$

To help to justify the results it is relevant to present the algorithm settings of each of the techniques used to build the classification models. In Table 3 is possible observe the available settings for four techniques: Support Vector Machine (SVM), Decision Tree (DT), Naïve Bayes (NB) and Generalized Linear Model (GLM). Even though four techniques were used the Generalized Linear Model (GLM) technique did not have customizable options.

Table 3 Techniques' Algorithm Settings

Support Vector Machine	
Kernel Function	Gaussian
Tolerance Value	0,001
Active Learning	Yes
Decision Tree	
Homogeneity Metric	Gini
Maximum Depth	7
Minimum Records in a Node	10
Minimum Percent of Records in a Node	0,05
Minimum Records for a Split	20
Minimum Percent of Records for a Split	0,1
Naive Bayes	
Pairwise Threshold	0
Singleton Threshold	0
Generalized Linear Model	
Reference Class name	1
Missing Value Treatment	Mean Mode
Ridge Regression	Disable
Approximate Computation	Disable

3.5 Evaluation

The evaluation phase analyzes the results given by all the created scenarios and concludes which have the best results. The results of the scenarios' performance and Receiver Operating Characteristic (ROC) curves were analyzed. Then a confusion matrix for each of the modeled scenarios was created. Using the confusion matrix were analyzed the results for each of the metrics: sensitivity, specificity and accuracy.

Each of these three metrics dictates a different prediction result. If we are trying to predict if a patient will need to take a vasopressor, the model with higher sensitivity should be selected. In case we are looking to predict if the patient would not need to take a vasopressor, the selected model should be the one with higher specificity. If both of the objectives are our target then a model with high accuracy should be selected.

Since in this case the objective is to get a model that will predict if a patient will need to take a vasopressor, a model with high values in the sensitivity metric is the one to look after.

4.0 DISCUSSION

The results analysis concluded that the models were very sensitive. This was the desired because sensitive models gives the true positive rate, i.e., predicts better if a patient will need to take a vasopressor.

In order to decide between the various models which is the best, threshold was established to each of the metrics:

- 85% to sensitivity
- 70% to accuracy
- 60% to specificity

In Table 4 it is possible to observe the three best models per metric that gave values above the defined limits.

Table 4 Top 3 Models per Metric

Sensitivity		
Scenario	**Technique**	**Value**
S1	SVM	90.72%
S1	NB	89.22%
S3	NB	88,68%
Specificity		
Scenario	**Technique**	**Value**
S3	SVM	82.72%
S4	SVM	76.00%
S1	NB	70.44%
Accuracy		
Scenario	**Technique**	**Value**
S3	SVM	86.02%
S4	SVM	82.84%
S1	SVM	81.14%

Having into account the established threshold for the confusion matrix's metrics, the results were very satisfactory. The best result was achieved in sensitivity metric with the model that had scenario S1 and technique SVM, presenting a result of 90.72%. In Figure 2 it is possible to observe the ROC curve of this model. This model has an area under curve of 83.06%, a max overall accuracy of 82.20%, a max average accuracy of 82.77% and a model accuracy of 77.96%.

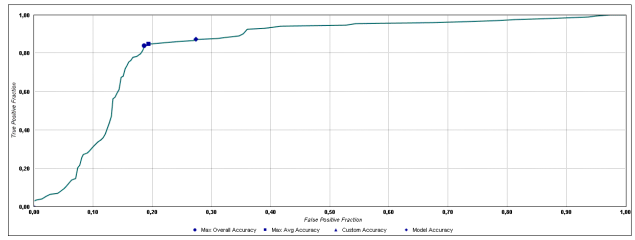


Figure 2 ROC Curve of the most sensitive model

In regard to the techniques, it is noticeable that Naïve Bayes and Support Vector Machines were the only ones capable of surpass the imposed limits. Also the presence of the Case Mix group proved to be necessary to acquire better results, since S2 did not even make it into the three most sensitive models. This means that non-clinical information was relevant to improve the results.

5.0 CONCLUSION

This study provides interesting results to the clinical science. The models induced can help the decision making process of intensivists in Intensive Care Units (ICU), by providing more information therefore allowing them to make more precise decisions. By predicting if a patient will need to take a vasopressor, intensivists will be able to act beforehand, making sure that the patient would not need to take the vasopressor when the predicted time comes. This benefits the institution as a whole, since helps reduce costs associated to the use of vasopressors and also the patients' health by preventing the need for vasopressors in the first place which sometimes can have side effects.

As future work, it is envisioned to reproduce these models but with the difference that the values from the Vital Signs and Lab Analysis groups will be grouped into classes with clinical meaning, else the use of statistical distribution could insert values into the wrong classes.

Acknowledgement

This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope UID/CEC/00319/2013 and PTDC/EEI-SII/1302/2012 (INTCare II)

References

[1] Direção Geral de Saúde. 2003. *Cuidados Intensivos: Recomendações para o seu desenvolvimento*. D. G. d. Saúde. Lisboa.

[2] Koh, H. C. and Tan, G. 2011. Data Mining Applications in Healthcare. *Journal Of Healthcare Information Management.* 19(65).

[3] Silva Á. J. B. M. d. 2007. Modelos De Intelegência Artificial Na Análise Da Monitorização De Eventos Clínicos Adversos, Disfusão/Falência De Orgãos E Prognóstico Do Doente Crítico.

[4] Ramon, J., Fierens, D., Güiza, F., Meyfroidt, G., Blockeel, H., Bruynooghe, M. *et al.* 2007. Mining Data From Intensive Care Patients. *Advanced Engineering Informatics.* 21: 243-256.

[5] Elliott, J. 1997. Alpha-Adrenoceptors In Equine Digital Veins: Evidence For The Presence of Both Alpha~ 1 and Alpha~ 2-Receptors Mediating Vasoconstriction. *Journal of Veterinary Pharmacology And Therapeutics.* 20: 308-317.

[6] Medscape. 2015, June 22. *Dopamine Dosing, Indications, Interactions, Adverse Effects And More*. Available: http://reference.medscape.com/drug/intropin-dopamine-342435.

[7] Medscape. 2015, June 22. *Epinephrine Dosing, Indications, Interactions, Adverse Effects And More*. Available: http://reference.medscape.com/drug/epipen-jr-epinephrine-342437.

[8] Medscape. 2015, June 22. *Norepinephrine Dosing, Indications, Interactions, Adverse Effects And More*. Available:http://reference.medscape.com/drug/levarterenol-levophed-norepinephrine-342443#0.

[9] American Heart Association. 2005. 7.2: Management of Cardiac Arrest. *Circulation.* 112: 60.

[10] Dellinger, R. P., Levy, M., M. Rhodes, A. Annane, D. Gerlach H., Opal, S. M. *et al.* 2013. Surviving Sepsis Campaign: International Guidelines for Management of Severe Sepsis And Septic Shock, 2012. *Intensive Care Medicine.* 39: 165-228.

[11] Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A. *et al.* 1992. Definitions for Sepsis and Organ Failure and Guidelines for The Use of Innovative Therapies In Sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine. *Chest Journal.* 101: 1644-1655.

[12] Portela, C. F., Santos, M. F., Silva, Á., Machado, J. and Abelha, A. 2011. Enabling a Pervasive Approach for Intelligent Decision Support in Critical Health Care. in *ENTERprise Information Systems*. Springer 2011. 221: 233-243.

[13] Portela, F., Santos, M., Machado, J., Abelha, A., Silva, Á., and Rua, F. 2014. Pervasive and Intelligent Decision Support in Intensive Medicine – The Complete Picture. in *Information Technology in Bio- and Medical Informatics*. Springer International Publishing. 8649: 87-102.

[14] Portela, F., Santos, M. F., and Vilas-Boas, M. 2012. A Pervasive Approach to a Real-Time Intelligent Decision Support System in Intensive Medicine. *Communications in Computer and Information Science*. Springer 2012. 0272: 14.

[15] Filipe Portela, Jorge Aguiar, Manuel Filipe Santos, Álvaro Silva, Fernado Rua. 2013. Pervasive Intelligent Decision Support System - Technology Acceptance in Intensive Care Units. *Advances in Intelligent Systems and Computing*. Springer.

[16] Portela, F., Santos, M. F., Machado, J., Abelha, A., and Silva, Á. 2013. Pervasive and Intelligent Decision Support in Critical Health Care Using Ensembles. in *Information Technology in Bio- and Medical Informatics*. Springer, 2013. 8060: 1-16.

[17] Braga, P., Portela, F., Santos M. F., and Rua F. 2014. Data Mining Models to Predict Patient's Readmission in Intensive Care Units. presented at the ICAART - *International Conference on Agents and Artificial Intelligence*.

[18] Portela F., Veloso R., Santos M. F., Machado J. M., Abelha A., Silva Á., *et al*. 2014. *Redict Hourly Patient Discharge Probability in Intensive Care Units using Data Mining*.

[19] Oliveira, S., Portela, F., and Santos, M. F. 2014. Pervasive Universal Gateway for Medical Devices. in *Recent Advances in Electrical Engineering and Education Technologies (SCI 2014)*. 205-210.

[20] Sérgio Oliveira, F. P., Manuel Filipe Santos, José Machado, António Abelha, Álvaro Silva and Fernando Rua. 2015. Intelligent Decision Support to predict patient Barotrauma risk in Intensive Care Units. *Procedia Technology - HCIST 2015 - Healthy and Secure People*. Elsevier.

[21] Filipe Portela S. O., Manuel Filipe Santos, António Abelha, José Machado. 2015. A Real-Time Intelligent System for tracking patient condition. *LNCS - Ambient Intelligence for Health*. Springer. 9456.

[22] Santos, M. F., Portela, F., and Vilas-Boas, M. 2011. *Intcare: Multi-Agent Approach For Real-Time Intelligent Decision Support In Intensive Medicine*.

[23] Filipe Portela, Filipe Pinto, and Santos M. F. 2012. Data Mining Predictive Models For Pervasive Intelligent Decision Support In Intensive Care Medicine. *KMIS* 2012. Barcelona.

[24] Santos, M. F. and Portela F. 2011. Enabling Ubiquitous Data Mining in Intensive Care - Features Selection and Data Pre-Processing. *ICEIS*. Beijing, China.

[25] Portela, F., Santos, M. F., Gago, P., Silva, Á., Rua, F., Abelha, A., *et al*. 2011. Enabling real-time intelligent decision support in intensive care". *ESM'2011*. Guimarães, 2011. 446.

[26] Hardin J. M. and Chhieng D. C. 2007. Data Mining and Clinical Decision Support Systems. in E. S. Berner, FACMI, and FHIMSS (eds.) *Clinical Decision Support Systems*. New York: Springer.

[27] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. 1996. From Data Mining to Knowledge Discovery in Databases. *AI magazine.*17: 37.

[28] Chen, M.-S., Han, J., and Yu, P. S. 1996. Data Mining: An Overview from a Database Perspective. *Knowledge and data Engineering, IEEE Transactions.* 8: 866-883.

[29] Harper, G. and Pickett, S. D. 2006. Methods for Mining HTS Data. *Drug Discovery Today.* 11: 694-699.