Predicting Plateau Pressure in Intensive Medicine for Ventilated Patients

Sérgio Oliveira[1], Filipe Portela[1], Manuel F. Santos[1], José Machado[1], António Abelha[1]
Álvaro Silva[2], Fernando Rua[2]
[1]Algoritmi Centre, University of Minho, Portugal
sergiomdcoliveira@gmail.com; {cfp, mfs}@dsi.uminho.pt
{jmac, abelha}@di.uminho.pt
moreirasilva@me.com; fernandorua.sci@chporto.min-saude.pt

Abstract Barotrauma is identified as one of the leading diseases in Ventilated Patients. This type of problem is most common in the Intensive Care Units. In order to prevent this problem the use of Data Mining (DM) can be useful for predicting their occurrence. The main goal is to predict the occurence of Barotrauma in order to support the health professionals taking necessary precautions. In a first step intensivists identified the Plateau Pressure values as a possible cause of Barotrauma. Through this study DM models (classification) where induced for predicting the Plateau Pressure class ($>=30$ cm$H_2$0) in a real environment and using real data. The present study explored and assessed the possibility of predicting the Plateau pressure class with high accuracies. The dataset used only contained data provided by the ventilators. The best models are able to predict the Plateau Pressure with an accuracy ranging from 95.52% to 98.71%.

Keywords: Barotrauma, Plateau Pressure, Intensive Medicine, Data Mining, INTCare, Mechanical Ventilation.

1 Introduction

With the advancement of technology, organizations have a set of data acquisition mechanisms able to collect information from various business processes. In this point organizations are recognized as a source of implicit knowledge. The process of extracting information from knowledge bases can provide better working practices or even redesign work processes [1]. Health institutions, in particular hospitals, possess multidisciplinary databases that can be an important knowledge source for predicting or diagnosing potential complications in patients. The development and application of Data Mining (DM) models has been experimented and proved to be very useful [2]. Using this new knowledge, physicians are able to identify more effective treatments and best practices. Consequently, patients can receive healthcare according to their condition. DM provides not only the methodology, but also the technology to transform the collected data into useful knowledge for the decision [3].

In the intensive medicine field one of the most respiratory diseases verified it is barotrauma[4]. It was identified by the physicians that high values of Plateau Pressure value [5] can contribute to a patient has barotrauma.

Framed in the INTCare project [8] and as a first step in order to avoid barotrauma it was explored DM techniques to predict (classification approach) the Plateau Pressure using as input only the data provided by the ventilator. This work used real data provided by the Intensive Care Unit of Centro Hospitalar do Porto – Hospital Santo António.

This paper consists of six chapters. In the first chapter is made a short introduction to the problem presented. Then presents the aspects related to the problem studied, the concepts and technology used in knowledge discovery in database (KDD) process. In the third chapter is presented the development of the practical component that had as a guideline the CRISP-DM methodology. The fourth chapter discusses the results and the most relevant aspects of the work. In the fifth chapter, relevant conclusions are presented. Finally, possible directions for future work are presented.

2 Background

2.1 Pressure Plateau and ARDS for Barotrauma

Barotrauma is the terminology used for identifying a patient with mechanical ventilation complications. The incidence of barotrauma associated with mechanical ventilation may concern the underlying disease for each patient. The occurrence of barotrauma is greater in patients with severe lung disease, especially patients with Acute Respiratory Distress Syndrome (ARDS) and pre-existence of chronic lung diseases [4].

Plateau pressure or static elastic recoil pressure of the respiratory system is useful to evaluate the elastic properties and the respiratory system resistance of patients with respiratory failure. Plateau Pressure (PLP) provides important information for the diagnosis of patients and should be monitored consecutively. It is important to maintain PLP values $<=$ 30 cmH_2O in order to protect the lungs. The increase of PLP values is associated with increased respiratory elastic system and with the decreasing of the respiratory system compliance and vice-versa. [5].

Although mechanical ventilation is able to support and assist the continuity of patient's life, mechanical ventilation can be harmful to the lungs. Patients such has ARDS the incidence of pneumothorax and barotrauma varies between 0% and 76%. It has been demonstrated through a study using a database of patients with ARDS, that there was not relationship between the ventilator settings and the occurrence of air leak [6]. The occurrence of barotrauma is increased in patients with pulmonary illnesses, especially in patients with ARDS, pneumonia and obstructive pulmonary disease. Researchers argue that there is a relationship between Positive End-Expiratory Pressure (PEEP) and the occurrence of barotrauma, however there are researchers which did not detect any relationship between PEEP and barotrauma.

Amato et al [4] argue that tidal volume of 12 ml/kg and high plateau pressures results in 42% rate of pneumothorax and a mortality rate about 71%.

2.2 Related work

Some researches had been made in order to understand the relationship between the Plateau pressure and the occurrence of barotrauma.

One study was designed to demonstrate how high pressure could increase the risk of barotrauma in patients with Acute Respiratory Distress Syndrome (ARDS). The respective study [6] verified that in ventilated patients with ARDS the occurrence of barotrauma is unusual these days. The occurrence of barotrauma may vary with the severity of lung disease, but the standard mechanical ventilation could explain the occurrence of issues, especially when Plateau Pressure is above 35 cmH_2O [6]. Also the incidence of Mediastinal Emphysema (ME) and air Pneumothorax were analysed to identify radiographic patterns and risk factors for the occurrence of Barotrauma.

Another study was conducted in a population of patients receiving mechanical ventilation in an intensive care unit. The study found that the peak values of inspiratory pressure, the pressure level at the end of expiration, the respiratory rate and the tidal volume were significantly higher in patients who developed barotrauma when compared with the other patients. High values reflect the high incidence of barotrauma in patients with ARDS [7].

2.3 INTCare

This study is being developed under the research project INTCare. INTCare is an Intelligent Decision Support System (IDSS) [8], which is in constantly developing and testing. It is deployed in the ICU of the Hospital Santo António – Centro Hospitalar do Porto (CHP).

The system monitors patient condition and uses data mining techniques for predicting patient outcome, patient organ failure, readmissions, length of stay, suggest procedures, treatments and therapies. The forecasting of Barotrauma is also one of its objectives. This work is the first step to predict Barotrauma.

2.4 Data Mining

For technical aspects, DM is a process that uses artificial intelligence techniques, statistics and mathematics to extract useful information and knowledge (or patterns) from large volumes of data. The discovery of patterns in the data may be in the form of business rules, affinities, correlations, or terms of prediction models [9]. For this work, the application of DM was achieved by using a statistical environment R.

R is presented as a programming language and environment for statistical development [10]. The e1071 library [11] was used to implement the techniques Support Vector Machine (SVM), Decision Tree (DT) and Naive Bayes (NB). To perform the evaluations of DM models it was used the rminer library [12].

3 Knowledge Discovering Process

The process of knowledge discovery using Data Mining techniques is complex, in that sense, Cross Industry Standard Process for Data Mining (CRISP-DM) methodology was followed to carry out the present study.

3.1 Business Understanding

This is the first phase of the CRISP-DM methodology. In this phase it was analysed the problem and defined the project requirements and its main goals. First, there is a need for developing this work because the ICU of CHP deals with mechanical ventilated patients and do not have any mechanism to prevent Barotrauma. Being the plateau pressure an important indicator and having in consideration all the evidences above mentioned, this consitutes the main reason to develop this work. In this context, the Data Mining goal was definied as: to predict and identify the variables (only monitored by ventilators) which are most correlate with the Plateau pressure. The data used for performing the present study were derived from the above-mentioned hospital. The DM models were induced by using a set of variables collected by the ventilators. The target is the variable corresponding to the Plateau pressure, which is a numeric variable grouped in two classes of values (<30 cmH_2O and $>=30$ cmH_2O).

3.2 Data Understanding

The initial data sample contains several records without patient identification (PID), i.e., the data collected does not correspond to anyone. This happened bacause sometimes the patients are admitted (for a few hours) in the ICU but is not assigned to Electronic Health Record. These records were discarded from this study. The sample used was collected from the ventilators and comprises the period between 01.09.2014 and 11.07.2014 and a total of 18 377 records. Each record contais sixteen fields:

- Date: corresponds to the day, month and year of the value collected by the ventilator;
- Bed: is the bed number where the patient was being ventilated;
- CDYN – (F_1): Dynamic compliance in mL/ cmH_2O;
- CSTAT – (F_2): Static compliance from inspiratory pause measured in mL/ cmH_2O;
- FIO2 – (F_3): Fraction of inspired oxygen (%);
- Flow – (F_4): Peak flow setting in litters per minute;
- RR – (F_5): Respiratory rate setting in berths per minute;
- PEEP – (F_6): Positive End-Expiratory Pressure in cmH_2O;
- PMVA – (F_7): Mean airway pressure in cmH_2O;
- Plateau pressure – (F_8): End inspiratory in cmH_2O;

- Peak pressure – (F_9): Maximum circuit pressure in cmH_2O;
- RDYN – (F_10): Dynamic resistance in cmH_2O;
- RSTAT – (F_11): Static resistance from inspiratory pause measured in cmH_2O/L/s;
- Volume EXP – (F_12): Exhaled tidal volume in litters;
- Volume INS – (F_13): Tidal volume settings in litters;
- Volume Minute – (F_14): Exhaled minute volume litters;
- Ventilation Mode – (F_15): Ventilation mode (CMV, SIMV and CPAP),
- Process: It is the patient identification;

Table 1 presents a statistical analysis for each field collected from the ventilators. For each one of the numeric values it was analysed the minimum and maximum value collected, their average, standard deviation (stDev) and coefficient of variation.

Table 1 – Distribution of variables

	Max	Min	Average	stDev	Variation
CDYN	200	0	32.30	21.97	68.02%
CSTAT	126	0	17.80	19.95	112.04%
FIO2	100	22	57.10	15.05	25.95%
Flow	80	0	37.91	15.96	42.10%
RR	29	0	12.21	8.71	71.34%
PEEP	10	3	5.92	1.51	25.58%
PMVA	32	0	12.41	2.83	22.84%
Peak Pressure	85	0	30.87	10.28	33.31%
RDYN	100	0	16.35	7.37	45.08%
RSTAT	51	0	10.40	11.04	106.12%
Volume EXP	1.98	0	0.48	0.12	25.80%
Volume INS	0.71	0	0.42	0.16	39.36%
Volume Minute	30.1	0	9.18	2.05	22.32%

Through Table 1 it is possible to verify the heterogeneity of each variable. Additionally, two more variables (not quantitative) were considered: Plateau Pressure (F_8) and the ventilation Mode (F_15). Plateau Pressure has only two possible classes ("<30 cmH_2O " and ">=30 cmH_2O "), distributed in a proportion of 85.92% for the class"<30 cmH_2O " and the remaining 14.02% for the class "> = 30 cmH_2O ". Ventilation Mode (F_15) has three possible classes ("CMV", "SIMV", and "CPAP"). Approximately 65.14% of the records have the "CMV" result, 4.87% of the records have the SIMV result and 29.94% of the records have the "CPAP" result.

3.3 Data Preparation

Since the present study is to make predictions using data mining techniques (classification approach) it was necessary to make some transformations in the data collected. First, some changes were made in the data source in order to transpose

values from the rows to the columns. Then, the records that had at least one null value were removed. Finally, also were made changes to the target variable, because the goal is to predict a Plateau Pressure class and this pressure is a quantitative variable. This variable was divided into two classes, according to scientific studies and ICU physicians. Values less than 30 cmH_2O are classified as normal.

3.4 Modelling

The Data mining techniques used to induce classification models were: Support Vector Machine (SVM), Decision Trees (DT) and Naive Bayes (NB). The choice of these techniques was based on two characteristics: interpretability and efficiency. The SVM reaches the second characteristic, but the DTs and NBs meet the two characteristics. To implement evaluation mechanisms and to test the induced model it was applied 10 Folds Cross Validation (10-Folds CV). The 10-Folds CV was adopted due to the good results demonstrated in multidisciplinary data [13]. All technical underwent tuning function. This feature comes with the e1071 package. The main objective is to perform research network ranges from hyper parameters previously provided and sequentially identify the best model and their hyper parameters.

The use of SVM technique is based on the application of two kernels: Linear and Radial-Basic Function. The two kernels handle different parameterizations because the hyperparameters are different for each kernel. Depending on the kernel used by SVMs, a range of values for parameter C was defined. Its range has been defined by the values obtained by the power $2^{(1,..,4)} = [2, ... ,16]$, where $C > 0$. The cost parameter C introduces some flexibility separating the categories in order to control the trade-off between errors in training or stiffness margins [14]. The hyper parameter Gamma (γ) was defined in the same way as C. The range was determined according to the values obtained by the power $2^{(-1,0,1)} = [0.5, 1, 2]$. Its parameterization was used in the RBF kernel. The γ value determines the curvature of the boundary decision [15].

The application of the DT technique was achieved by CART algorithm. The feature selection methods and rules of decomposition were applied: Information Gain (IG) and the Gini Index (GI). The attribute selection measure IG determines the attribute with the highest information gain and uses it to make the division of a node [16]. The GI is determined by the difference between the original information requirement (i.e., based on only the ratio of classes) and the new requirement (i.e., obtained after partitioning A). The respective difference can be demonstrated as follows:$Gain(A) = Info(D) - Info_A(D)$. The attribute A has the highest information gain. Gain (A) is the division attribute of node n [16]. The objective of GI it is to calculate the value for each attribute using the attribute for the node with the lowest impurity index [1]. The GI index measures the impurity of D, using a data partition or a training set of attributes $Gini(D) = 1 - \sum_{i=1}^{m} pi^2$, where pi corresponds to the probability of an attribute D of a class Ci. This value is estimated by $|Ci \cap D|/|D|$. The sum is calculated as a function of m classes [16]. Finally, in NB algorithm there was not any configuration, but as already described earlierthis

algoritm uses the tune function to identify the sampling method to be used. All the configuration was previously determined.

The developed models can be represented by the following expression:

$$M_n = \{A_f;\ F_i;\ D_x;\ TDM_y;\ MA_k\}.$$

The model M_n belongs to an approach (A) and it is composed by Fields (F), a type of variable (TV) a DM technique (TDM) and a sampling method (SM):

$$A_f = \{Classification_1\}$$
$$F_i = \{V\_1_1, V\_2_2, V\_3_3, V\_4_4, V\_5_5, V\_6_6, V\_7_7, V\_8_8, V\_9_9, V\_10_{10}, V\_11_{11}$$
$$, V\_12_{12}, V\_13_{13}, V\_14_{14}\}$$
$$TV_x = \{Qualitative\ variables\ ordinal_1\}$$
$$TDM_y = \{SVML_1, SVMRBF_2, DTGI_3, DTIG_5, NB_6\}$$
$$SM_k = \{10 - folds\ CV_1\}$$

Through the notation for representing DM models it was possible to present an example of an implemented model. A DM (M_1) model following the classification approach, using the data from the fields (F_1, F_2, F_3, F_4, F_5, F_6, F_7, F_10 and F_11), the SVM technique with RBF kernel and the sampling method 10-folds CV is expressed by:

$$M_1 = A_1 + F_{\{1,2,3,4,5,6,7,10,11\}} + TV_1 + TDM_2 + SM_1.$$

3.5 Evaluation

In order to evaluate the results presented by the DM models three metrics were considered: Accuracy, Specificity and Sensitivity.

The dataset used in the training phase it was divided in exclusive subsets through the 10-folds CV. In the implementation of the respective dividing procedures, ten executions were performed for each one of them. About 100 experiments were performed for each test. Table 2 presents the four best models obtained and their metrics: Accuracy (ACC), Sensitivity (SEN) and Specificity (SPE).

Table 2 - Valuation models

Model	Fields	TDM	SM	ACC	SEN	SPE
BM_1	$F_{\{1,2,3,4,5,6,7,8,9,\ 11,12,13,14,15\}}$	TDM_1	SM_1	98.71%	99.33%	94.95%
BM_2	$F_{\{1,7,9\}}$	TDM_1	SM_1	95.52%	98.28%	78.72%
BM_3	$F_{\{1,2,7,9\}}$	TDM_1	SM_1	96.54%	98.02%	87.55%
BM_4	$F_{\{1,2,7,9,11\}}$	TDM_1	SM_1	96.77%	98.83%	87.74%

The technique which provided the best result was the SVM with Linear kernel. However the kernel also was extremely useful for decision trees to demonstrate which fields had greater importance/relevance to the creation of the models. Naive

Bayes techniques did not show much relevance in these models because they were always lower than the results obtained by SVM, although the difference in accuracy measue was little bit noticeable (3% to 7%).

Decision Trees were very important to identify the input variables that influence the models. The set of variables with most influence is: CDYN ($F_{\{1\}}$), CSTAT ($F_{\{2\}}$), PMVA ($F_{\{7\}}$) and Peak pressure ($F_{\{9\}}$).

4 Discussion

The predictions made by the DM models were very satisfactory attaining accuracies between 95.52% and 98.71%, presenting also very good results at level of sensitivity (between 98.02% and 99.33). Thus the generated models are able to predict patients with Plateau pressure in the class "<30 cmH_2O " properly and the same applies to patients who have had the result of prediction "$> = 30$ cmH_2O ". Table 3 presents the Confusion Matrix for the best model:BM_1.

Table 3 – Confusion Matrix

Target	Predictive	
	<30 cm cmH_2O	**>=30 cm** cmH_2O
<30 cm cmH_2O	15677	131
>=30 cm cmH_2O	106	2463

Through Table 3 it is possible to identify that there is a large imbalance between the classes. Approximately 86.02% of the predictions made correspond the class "<30 cm cmH_2O" and about 13.98% presented as result the class "$> = 30$ cmH_2O". However the BM_1 can predicts well as "<30 cmH_2O" and "$> = 30$ cmH_2O".

Scenarios (Fields column) shown in Table 2 were the ones who gave the best results. BM_3 model used only 4 inputs and obtained 96.55% of accuracy. BM_4 model used the same inputs of BM_3 plus one, but the outcome of the model is only better in 0.24%. The model BM_3 only used the variables with greatest importance.

Decision trees were important to identify the most important variables. Naive Bayes did not represent an important technique for this study. Also should be noted that the Support Vector Machine were the technique which presented better results. Linear kernel provided the best results.

5 Conclusion

This study explored the prediction of Plateau pressure class using data only provided by the ventilators. The best results obtained in terms of accuracy vary between 95.52% and 98.71%. Although there is wide variation in the number of occurrences of each class. It was possible to demonstrate through the results achieved by the accuracy metric that the best model generated BM_1 did not have great difficulty in predicting both classes. On the other hand the other models do not demonstrate the same performance. They revealed more difficulty when

predicting the class ">= 30 cmH_2O ". Despite of these results all the models are using real data and are executed in real-time meaning that the models may present different performances with new data.

Support Vector Machine obtained the best results. Decision Trees were important to identify the variables that most influence the models.

The models presented demonstrate their viability predicting the Plateau Pressure class and consequently constitute an important help to prevent the barotrauma occurrence.

6 Future Work

For future research will take into consideration the following aspects:
- To consider more variables such as gender, age, respiratory disease, if the patient is a smoker and others. These variables must be added to the prediction models in order to improve accuracy;
- To explore the regression approach to predict continuous values;
- To make predictions by hour for each one of the ventilated patients;
- To embed these models into the INTCare Decision Support module and identify whether the predictions made are valid.

Acknowledgements

References

[1] M. Santos and C. Azevedo, *Data Mining Descoberta do conhecimento em base de dados*. FCA - Editora de Informática, Lda, 2005.

[2] M. Santos, M. Boa, F. Portela, Á. Silva, and F. Rua, "Real-time prediction of organ failure and outcome in intensive medicine," in *2010 5th Iberian Conference on Information Systems and Technologies (CISTI)*, 2010, pp. 1 –6.

[3] H. Koh and G. Tan, "Data mining applications in healthcare," *J Healthc Inf Manag*, vol. 19, no. 2, pp. 64–72, 2005.

[4] A. Anzueto, F. Frutos-Vivar, A. Esteban, I. Alía, L. Brochard, T. Stewart, S. Benito, M. J. Tobin, J. Elizalde, F. Palizas, C. M. David, J. Pimentel, M. González, L. Soto, G. D'Empaire, and P. Pelosi, "Incidence, risk factors and outcome of barotrauma in mechanically ventilated patients," *Intensive Care Med*, vol. 30, no. 4, pp. 612–619, Apr. 2004.

[5] N. Al-Rawas, M. J. Banner, N. R. Euliano, C. G. Tams, J. Brown, A. D. Martin, and A. Gabrielli, "Expiratory time constant for determinations of plateau pressure,

respiratory system compliance, and total resistance," *Crit Care*, vol. 17, no. 1, p. R23, 2013.

[6]   M. Boussarsar, G. Thierry, S. Jaber, F. Roudot-Thoraval, F. Lemaire, and L. Brochard, "Relationship between ventilatory settings and barotrauma in the acute respiratory distress syndrome," *Intensive Care Med*, vol. 28, no. 4, pp. 406–413, Apr. 2002.

[7]   R. B. Gammon, M. S. Shin, and S. E. Buchalter, "Pulmonary barotrauma in mechanical ventilation. Patterns and risk factors," *Chest*, vol. 102, no. 2, pp. 568–572, Aug. 1992.

[8]   F. Portela, M. F. Santos, J. Machado, A. Abelha, Á. Silva, and F. Rua, "Pervasive and Intelligent Decision Support in Intensive Medicine – The Complete Picture," in *Information Technology in Bio- and Medical Informatics*, M. Bursa, S. Khuri, and M. E. Renda, Eds. Springer International Publishing, 2014, pp. 87–102.

[9]   E. Turban, R. Sharda, and D. Delen, *Decision Support and Business Intelligence Systems*, 9ª Edição. Prentice Hall, 2011.

[10]   L. Torgo, *Data Mining with R: Learning with Case Studies*. CRC Press - Taylor & Francis Group, 2011.

[11]   D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch, "Misc Functions of the Department of Statistics (e1071)." 2012.

[12]   P. Cortez, "Simpler use of data mining methods (e.g. NN and SVM) in classification and regression." 2013.

[13]   I. Witten, E. Frank, and M. Hall, *Data Mining Pratical Machine Learning Tools and Techniques*, 3ª Edição. Morgan Kaufmann, 2011.

[14]   M. Kantardzic, *Data Mining Concepts, Models, Methods, and Algorithms*, 2ª Edição. Wiley - IEEE Press, 2011.

[15]   A. Ben-Hur and J. Weston, "A User's Guide to Support Vector Machines," in *Data Mining Techniques for the Life Sciences*, O. Carugo and F. Eisenhaber, Eds. Humana Press, 2010.

[16]   J. Han, M. Kamber, and J. Pei, *Data Mining Concepts and Techniques*, 3ª Edição. Morgan Kaufmann, 2012.