Proceedings of the 18th World Congress
The International Federation of Automatic Control
Milano (Italy) August 28 - September 2, 2011

Preprocessing of Clinical Databases to improve classification accuracy of patient diagnosis*

**Filipe J. Marques*, Alexandra Moutinho*,
Susana M. Vieira*, João M. C. Sousa***

*Technical University of Lisbon, Instituto Superior Técnico, Dept of Mechanical Engineering, CIS/IDMEC – LAETA, Av. Rovisco Pais, 1049-001 Lisbon, Portugal (e-mail: filipe_jose_mmm@hotmail.com, alexandra.moutinho@ist.utl.pt, susana.vieira@ist.utl.pt, jmsousa@ist.utl.pt).

Abstract: In this paper, the prime importance of preprocessing in clinical databases is discussed. Specifically in intensive care units, data is often irregularly recorded, contain a large amount of missing values and sampling times are uneven. This paper proposes a systematic preprocessing procedure that can be generalized to common clinical databases. This procedure is applied to a known septic shock patient database and classification results are compared with previous studies. The goal is to estimate, as accurately as possible, the outcome (survived or deceased) of these septic shock patients. Neural modeling is used for classification. Detailed classification results are presented and show that the preprocessing is crucial to improve classifiers accuracy.

Keywords: Preprocessing, Clinical database, Neural network, Medical applications, Sepsis

1. INTRODUCTION

Nowadays, databases are typically very large and contain a high percentage of missing data. In most cases, the missing data come from multiple heterogeneous sources (Han et al., 2006). After data collection and problem definition, preprocessing is very important for data analysis, especially for retrospective evaluations. Medical databases are a good example where the preprocessing is essential. Clearly, the quality of the results from data analysis strongly depends on the careful execution of the preprocessing step (Brause et al., 2001).

In an intensive care unit (ICU), patients in an advanced stage of sepsis carries a high burden, namely a high mortality rate (about 50%) and higher costs of treatment compared with other ICU patients (Paetz, et al., 2002). After an operation, patients have a tendency to develop a phenomenon related to the mechanism of immune system. The pathophysiology of sepsis in humans is poorly understood. This syndrome was defined by consensus statement in 1992 to consist of certain criteria. Severe sepsis was defined as organ failure in the setting of sepsis, and septic shock was defined as severe sepsis where the organ failure was hypotension (Warren, 2009).

Over the past years knowledge-based neural networks and neuro-fuzzy techniques have been applied in the domain of outcome prediction for septic shock patient (Paetz, 2003). They stated that the preprocessing is a very important step for analysis of medical data analysis. They suggest that preprocessing should be an interdisciplinary work from data analysts and physicians. The quality of the database strongly depends on the success of data collection. Due to many factors it is almost impossible to get a 100% clean database of different patient records (Paetz, et al., 2000).

Our main goal is to devise a systematic preprocessing procedure applicable to clinical databases in order to obtain a reasonable quality of data. The proposed preprocessing approach was applied to a septic shock patient database, and the results obtained were compared with the ones presented in Paetz (2003). The results from the data analysis show that the careful execution of the preprocessing steps strongly influence the accuracy of the obtained classifiers. A good preprocessing can reduce model's complexity. In this paper we show the fundamental idea of preprocessing, together with neural network modeling, being advantageous in the medical domain.

The paper is organized as follows. Section 2 describes the proposed preprocessing procedure, i.e., to what kind of clinical databases can this preprocessing procedure be applied, the most common problems found in a clinical database and a description of the preprocessing procedure. In Section 3 the neural network model is presented. The preprocessing algorithm and neural network model are applied to a septic shock patient database and the obtained results are presented in Section 4. Finally, conclusions are drawn in Section 5.

* This work was supported by Fundação para a Ciência e a Tecnologia, through IDMEC under LAETA, project PTDC/SEM-ENR/100063/2008, and by a FCT grant SFRH/BPD/65215/2009, Fundação para a Ciência e a Tecnologia (FCT), Ministério do Ensino Superior, da Ciência e da Tecnologia, Portugal.

2. PREPROCESSING OF CLINICAL DATABASES

Preprocessing is a very important step in data analysis and can heavily influence the modelling results. In the following, we will show the main problems associated with clinical databases. According to Brause et al (2001), a large amount of missing data and uneven sampling times are typical for medical data and should be taken into account in all approaches for medical data diagnosis. To get a good clean data from a medical database, a systematic preprocessing procedure is proposed and is described in more detail in this section.

2.1 Most Common Problems in Clinical Databases

Typically, the data is collected by several hospitals. As such, in each hospital different variables are measured. Naturally, the medical retrospective data material is very heterogeneous, a fact that has to be emphasized (Paetz et al. 2003). The typical problems associated with medical databases are listed below:

1. Each of the patients has a different length of stay in the medical unit.
2. For each patient, a different number of variables is documented.
3. Different data is measured at different times of day with a different frequency.
4. Some hospitals may not have data recorded online. Since the data can be transferred from handwritten records to the database, typing errors are a common error source.
5. Many variables have a high percentage of missing values caused by faults or simply by seldom measurements.

Nevertheless, with the preprocessing approach proposed in this work, it was possible to achieve better modelling results.

2.2 Preprocessing Procedure for Clinical Databases

With a proper systematic preprocessing procedure it is possible to improve the quality of the results. In figure 1 the scheme of the proposed preprocessing procedure is presented. In the following, this preprocessing procedure is explained step by step.

1. This first step helps to select patients and variables that have a small percentage of missing data, which guarantees some quality of the information used for modelling. The selection of the patients and variables is based on four types of analysis:
   a. Calculate the average of the sampling periods of variables. This analysis helps to find the variables that have a sampling period on average significantly bigger than 24 hours. These variables are discarded.
   b. Calculate the percentage of available data for each variable for the sampling times 1h, 12h and 24h. Knowing the day of entry and exit of the patient from the hospital it is possible to predict the number of sampled data needed for each sampling time of the patient. Knowing the amount of data that each patient has, the percentage of missing data per patient is estimated and can be visualized using an histogram.
   c. Determine the number of patients where the variable is measured. This step helps to choose the variables that are present in a large percentage of patients.

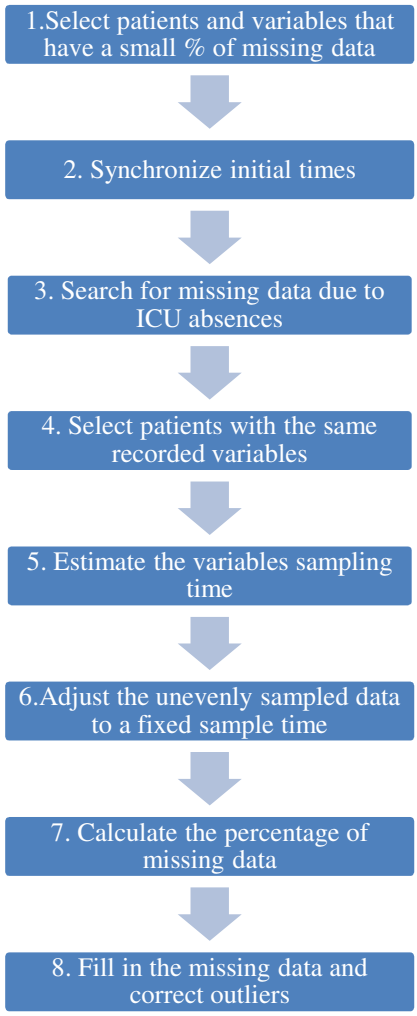

Fig. 1. Preprocessing procedure for clinical databases.

2. Normally, the time records of the data is saved in the time format: Day/Month/Year Hour:Minute. This time format is not ideal for the preprocessing and visualization of data in temporal graph. The entries of the patients are in different days, so this second step synchronizes the initial recording times. The initial time (t_0=0) is the same for all patients. This second step consists of two parts:
   a. Convert the time format to hours.
   b. Find the smallest hour of all time patient data and subtract the hour to all time of the patient.
3. Sometimes a patient can be absent from the hospital or just the ICU for a period of time. This lack of recorded

data can cause a big percentage of missing data. So the third step consists of finding missing data on variables in the same period of the day for each patient. Once detected the absence of the patient, only the data after the absence is considered.

4. To obtain a neural model it is required that patients have the same variables. As such, this step is very important to determine which variables and patients will be considered to train the model. This step consists of two parts:

   a. Obtain a binary matrix where rows are patients and columns are variables. If the patient X does not have data of variable Y, the cell of matrix (X,Y) is 0 (zero), otherwise is 1 (one).

   b. Eliminate rows and columns which contain 0 values, until the final matrix is unitary. The final result indicates which patients and variables are chosen for the next step and guaranty that all patients have the same variables.

5. After selecting the variables it is necessary to estimate the sampling time. This fifth step starts by determine and store the time differences between the sampled points of all patients for each variable. Then proceed to two types of analysis:

   a. Calculate the average and standard deviation of the time differences for each variable.

   b. Obtain an histogram of the time differences for each variable.

   From these two reviews it is possible to choose the best sampling time for each variable.

6. After defining the sampling times of each variable, the next step is to adjust the unevenly sampled data to the fixed sample time defined. The sampling times most used are 1h, 12h and 24h. Only the variables that have a sampling time smaller or equal to the chosen sample time, are adjusted.

   In figure 2 the adjustment of sampled data is explained. The t_n is time, spaced by a defined sampling time t_a, where the data should be adjusted. As shown in figure 2-a only the sampled data that are closer to time t_n, will be adjusted. In the case represented in figure 2-b, if the sampled data coincides with time t_n, then the data is not moved. If the sampled data is equally placed between two time instances, it will be set to one of the two time instances with equal probability, as shown in figure 2-c. The data that was not possible to adjust to a specific time instance will be discarded.

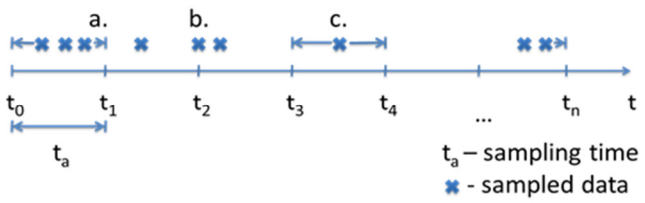

Fig. 2. Procedure of adjustment sampled data.

After this step, all variables are adjusted for the same time for each patient, but there is still missing data to fill in the next step.

7. After adjusting the sampled points it is possible to calculate accurately the percentage of missing data. For a better visualization of the results we used two types of histograms:

   a. Missing data rate per patient

   b. Missing data rate per variable.

8. The last step is to fill in missing data and correct outliers. There are many methods to fill in missing data. In this study the method used was ZOH (zero-order-hold). This method is simple and consists in filling missing data with the previous data value. In medical terms, this method approaches to the case when a doctor visits a patient and observes the last values stored in the clinical record. In this paper, outliers are defined as values that are outside the physiological limits of the variables. The outliers can occur due to typing errors, for example. For these cases the outlier is considered as missing data and it is replaced the same way as a given missing data, using ZOH.

3. NEURAL NETWORK MODELLING

In the last years many authors contributed to machine learning, data mining, intelligent data analysis and neural networks in medicine, e.g. Brause et al (2000) and Lavrac (1999). For our problem of septic shock diagnosis artificial intelligence techniques have the advantages of nonlinear classification, learning from data and good generalization ability. It is widely accepted in the medical community that septic shock dynamics are strictly nonlinear (Toweill et al, 2000). Our aim is to detect if a patient will survive or decease using a neural classifier.

There are many different types of neural networks and learning algorithms. Several optimization methods were tested, and the Levenberg-Marquardt backpropagation method provided the best resutls. The backpropagation algorithm performs learning on a multi-layer feed-forward neural network. A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer. Each layer is made up of units called artificial neurons. The untis are joined with varying connection strengths and weights. Each connection has an associated weight, which determines the effect of the incoming input on the activation level of the unit. The neuron output signal is given by the following relationship:

$$f(w^T, x) = f\left(\sum_{j=1}^{N} w_j x_j\right)$$

Where, for classification, $f(w^T, x)$ is a nonlinear activation function, $w = (w_1, ..., w_n)^T \in \Re^n$ is the weight vector, and $x = (x_1, ..., x_n)^T \in \Re^n$ is the vector of neuron inputs. For each training epoch, the weights are modified in order to minimize the mean squared error between the network's

prediction and the actual target value. The training stop criteria were an error of 0,001 or 300 epochs. The network approximates the underlying function, thus turning it into an optimization problem.

4. RESULTS

4.1 Medan database

This paper uses the public available Medan database (Hanisch, 2003), which contains the data 410 patients with abdominal septic shock. The data were recorded from 71 German intensive care units from 1998 to 2002 by medical documentation staff. The Medan database was originally composed by several tables, from which two were selected: patient information and variables measurements. The table of the variable measurements contains 103 different variables. The data is recorded as time series containing the date and time of recording.

4.2 Preprocessing of the Medan database

This section presents the results of the preprocessing procedure described in Section 2.2. The selection of variables with a small number of missing data leaded to the 103 variables presented in Table 5, see Appendix A, The different data are measured at different times of the day with a different frequency. It can also be observed in Table 5 that for each variable, a different number of patients is documented.

Figure 3 presents the percentage of missing data per patient. It is clear that the percentage of missing data is very high; around 50%. The distribution ressembles a normal distribution with a median value close to 50%.

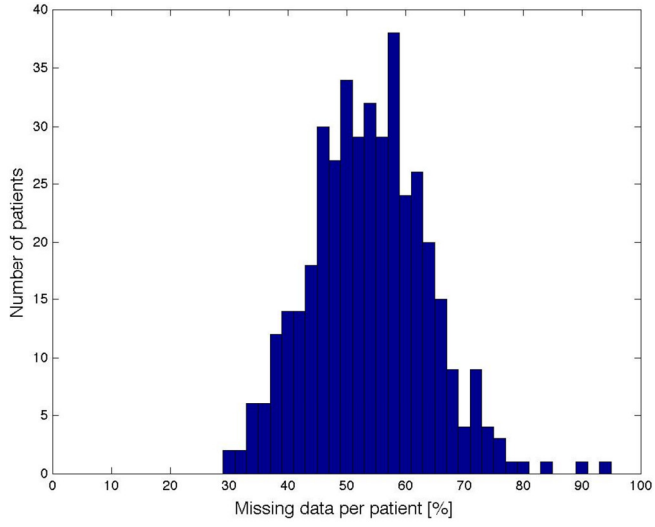

Fig. 3: Missing data per patient.

By analyzing Table 5, it was concluded that the variables with less than 250 patients will not be considered for modeling due to lack of information. Further, the average of sampling periods should be near or below 24 hours. After these steps, only 59 variables were considered to have relevant information.

The Step 3 of the algorithm proposed in Section 2.2 is important to find patients that have been away from an ICU for a long period of time. We found 10 patients who were absent for a period exceeding 24 hours. Figure 4 is an example of a patient that was in the ICU for more than 20 days. In these cases, we only considered the last part of the data, after their absence.

After Step 4, the variables and patients with relevant information for modeling are chosen. At this step we had 120 patients and 27 variables.

The sampling time for each variable is defined in Step 5. Figure 5 is an example of the sampling periods of the variable leukocytes. In this case, the sampling time was chosen to be 24 hours, as most sampling periods are indeed approximately 24 hours.

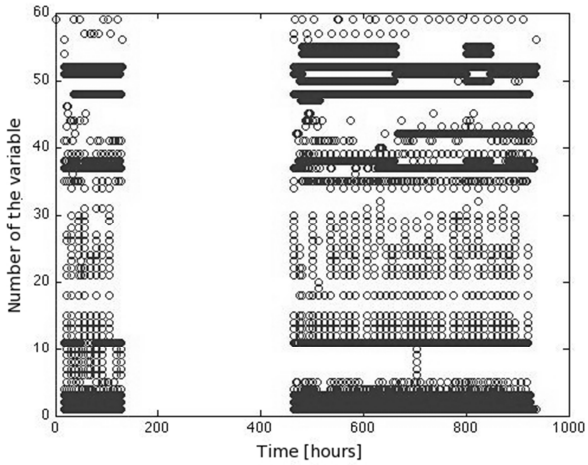

Figure 4: The absence of a patient in ICU for a period exceeding one day.

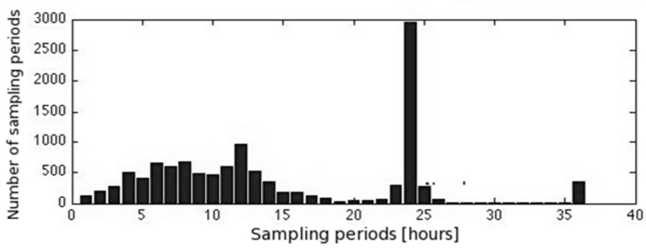

Fig. 5. Sampling periods of the variable leukocytes

Finally, the data were adjusted for the sampling time of 24 hours. The missing data were filled and the ouliers corrected with ZOH (zero-order-hold).

4.3 Predicting outcomes of sepsis patients

Recall that the goal of this paper is to apply a preprocessing procedure to a known septic shock patient database in order to predict sepsis patients' outcomes.

Two models were identified using neural networks, one using 12 variables (Model 1) and another using the 27 variables

(Model 2) obtained by the preprocessing procedure described in this paper. The model using 12 variables was derived in order to compare the outcome prediction capabilities of our models with the ones developed in Paetz (2003). Therefore, we use the same 12 variables as in Paetz (2003).

In order to evaluate the performance of the developed models, three different criteria were used: percentage of correct classification, sensitivity and specificity. Sensitivity measures the proportion of actual positives which are correctly identified as such (e. g. the percentage of sepsis patient who are correctly identified as having the condition deceased). Specificity measures the proportion of negatives that are correctly identified (e. g. the percentage of sepsis patient who are correctly identified as having the condition survived). For this particular application, the goal is to correctly classify which patients are more likely to decease, in order to rapidly act in their best interest. Bearing this in mind, the classifier should classify as accurately as possible the cases that result in death, or true positives, and aim to have a low number of false negatives. In other words, the classifier should maximize sensitivity.

Both models are the best neural network obtained after 10 simulations. Model 1 and Model 2 are identified using neural network with the following parameter: one hidden layer with 3 and 4 neurons, respectively, and in both the hidden layer and the output layer have, respectively, the transfer function 'logsig' and 'satlins'. The neural network was training by Levenberg-Marquardt backpropagation function. Training was done with 60% of the data, and the rest was used for testing, similar to the division of data used in Paetz (2003).

Table 2 presents the percentage of correct classifications for the developed models and for the model in Paetz (2003). It is clear that both Model 1, and especially Model 2, perform better than the one developed in Paetz (2003). Model 2 is about 10% better than the others, due to the larger amount of information given by the extra 15 variables.

Table 1. Percentage of correct classifications with mean value, standard deviation, minimum and maximum

		Correct classification			
Measure	**Number of variables**	**Mean value**	**Standard Deviation**	**Minimum**	**Maximum**
Paetz 2003	12	69	4,37	61,20	72,66
Model 1	12	70,51	2,28	67,35	74,49
Model 2	27	79,38	3,32	72.92	85,42

Table 3 and Table 4 show the results obtained for specificity and sensitivity, respectively. In terms of specificity, the model of Paetz (2003) has the best value. However, Model 2 has a value that is very close. Model 1 has a specificity that is slightly lower than the other two models. Recall that specificity predicts the patients that will survive.

The models developed using the preprocessing method described in this paper have much better values of sensitivity. Recall that this measure determines the patients that will decease. This is very important to hospital practitioners in order to begin the treatment of sepsis as fast as possible. By improving about 45%, our models are clearly much better.

Table 2. Percentage of specificity with mean value, standard deviation, minimum and maximum

		Specificity			
Measure	**Number of variables**	**Mean value**	**Standard Deviation**	**Minimum**	**Maximum**
Paetz 2003	12	92,26	-	-	-
Model 1	12	84,32	3,63	79,55	88,64
Model 2	27	91,72	5,91	86,21	100

Table 3. Percentage of sensitivity with mean value, standard deviation, minimum and maximum

		Sensitivity			
Measure	**Number of variables**	**Mean value**	**Standard Deviation**	**Minimum**	**Maximum**
Paetz 2003	12	15,01	-	-	-
Model 1	12	59,26	3,9	53,70	64,81
Model 2	27	60,53	6,2	52,63	68,42

In general, Model 2 with 27 variables presents clearly the best results. It has higher percentage of correct classifications (about 10% better), similar specificity and much better sensitivity (more than 45%). This fact validates the preprocessing procedure proposed in this paper.

5. CONCLUSIONS

This paper proposed a preprocessing procedure to be applied to medical databases, namely in this case a publicly available ICU database. Patients and variables were select using the preprocessing procedure, and neural models were derived base don the treated data. The proposed approach clearly outperformed a previous approach in terms of accuracy (percentage of correct classifications) and sensitivity, which is the most important measure for the application in hands.

In the future, this preprocessing procedure will be applied to larger health care databases which have more available features. To initially reduce the number of features, feature selection can be applied. Once, neural networks have good performance in terms of specificity and sensitivity, neural networks multi-models will be considered in future work based on work carried out in (Henriques et al., 2009).

REFERENCES

Brause, R., Hanisch, E. (2000), Medical Data Analysis ISMDA 2000. *Lecture Notes in Computer Science (LNCS)*, Volume 1933, Springer Verlag.

Brause, R., Hamker, F., Paetz, J. (2001). Septic shock diagnosis by neural networks and rule based systems. In: Jain, L. C. (Eds.), *Computational Intelligence Techniques in Medical Diagnosis and Prognosis*, pp. 323-356.

Han, J., Kamber, M. (2006), *Data Mining – Concepts and Techniques.* pp. 47, Morgan Kaufmann, San Francisco.

Hanisch, E., Brause, R., Arlt, B., Paetz, J., Holzer, K. (2003). The Medan Database, *http://www.medan.de* (accessed March 2010).

Henriques, J., Rocha, T., (2009). Prediction of Acute Hypotensive Episodes Using Neural Network Multi-models, *Computers in Cardiology,* Volume 36, pp. 549-552.

Lavrac, N. (1999), Machine Learning for Data Mining in Medicine. In: Horn, W. et al (Eds.), *Proc. AIMDM'99.* LNAI 1620. Springer-Verlag Berlin Heidelberg, pp. 47-62.

Paetz, J., Thone, F. (2000). About the Analysis of Septic Shock Patient Data*. Proceedings in the First International Symposium on Medical Data Analysis (ISMDA),* 1933*,* pp. 130-137.

Paetz, J., Arlt, B. (2002). A Neuro-Fuzzy Based Alarm System for Septic Shock Patients with a Comparison to Medical Scores. *Proceedings in the Third International Symposium on Medical Data Analysis (ISMDA),* Volume 2526*,* pp. 42-52.

Paetz, J., Arlt, B., Erz, K., Holzer, K., Brause, R., Hanisch, E. (2003). Data quality aspects of a database for abdominal septic shock patients. *Computer Methods and Programs in Biomedicine,* Volume 75, pp. 23-30.

Paetz, J. (2003). Knowledge-based approach to septic shock patient data using a neural network with trapezoidal activaton functions. *Artificial Intelligence in Medicine*, Volume 28, pp. 207-230.

Toweill, D., Sonnenthal, K., Kimberly, B., Lai, S., Goldstein, B. (2000). Linear and Nonlinear Analysis of Hemodynamic Signals During Sepsis and Septic Shock. *Critical Care Medicine,* Volume 28, pp. 2051-2057.

Warren, H. Shaw (2009), Editorial: Mouse Models To Study Sepsis Syndrome in Humans, *Journal of Leukocyte Biology,* Volume 86, pp. 199-201.

Appendix A.

Table 4. Average of sampling periods of the variables without any preprocessing or resampling, computed by data of all patients for which each variable was measured

Variable	Average (hours)	Standard deviation	Number of patients
Heart Rate	1,16	2,44	410
BPsys	1,17	2,13	410
BPdia	1,19	2,76	403
BPaverage	1,23	2,72	106
Temperature	2,13	4,01	410
CVP	4,39	8,02	402
PCWP	6,04	10,44	62
pH	6,01	14,46	389
PaO2	6,03	14,74	388
PaCO2	5,92	14,48	386
Base Excess	6,12	15,15	371
Bicarbonate	6,08	14,82	369
SpO2	1,94	6,780	400
Leukocytes	15,76	14,65	406
Erythrocytes	16,44	14,77	367
Haemoglobin	8,40	11,84	410
Haematocrit	11,01	13,22	406
Thrombocytes	15,80	13,80	393
TPZ	18,04	19,20	390
PTT	17,53	18,73	407
TT	24,68	37,99	216
AT3	23,28	45,30	348
Fibrinogen	25,19	40,79	313
erum Na+	9,22	13,29	410
Serum K+	8,51	12,37	410
Serum Ca++	19,68	34,73	356
Serum Cl-	21,37	38,62	249
Serum Creatinin	20,93	18,21	402
Urea	23,30	22,36	388
Uric Acid	60,41	117,05	131
GOT	38,95	47,18	386
GPT	40,08	47,98	391
GGT	43,15	57,34	354
Alcalic Phosphatase	44,79	57,56	330
Bilirubin Total	36,84	41,54	386
LDH	50,27	84,34	292
Cholesterine	73,23	105,72	199
Triglyceride	58,83	72,32	294
Albumin	45,30	54,94	226
Protein Total	37,05	52,81	274
CRP	28,16	23,87	371
Iron	45,94	56,44	44
Blood Sugar	5,83	9,812	391
Lung Infiltrate	58,67	85,43	266
Urine	1,43	3,73	396
Analgetica	1,81	9,22	409
Antiarrhythmic	4,11	24,66	220
Antibiotica	6,72	16,57	404
Antihypertensive	2,78	17,27	240
Antihypotensive	6,45	60,89	103
Anticoagulation	2,02	6,44	392
Antimycotic	11,04	17,61	131
Antipyretic	11,18	41,01	257
THAMTris	4,30	8,24	4
Sodium Bicarbonate	26,87	79,79	157
Broncholysis	2,89	17,34	124
Diuretica	18,55	37,50	135
Loop Diuretica	4,02	15,68	374
Immunoglobuline Polyglob	14,70	21,73	24
Pentaglobin	2,19	8,08	24
Immunosuppression	2,34	22,52	176
Crystalloides	4,88	16,90	394
Colloides	16,15	43,41	373
Erythrocytes Concentrate	39,73	73,11	341
FFP	19,47	58,58	265
Leukocytes Concentrate	0	0	0
Thrombocytes Concentrate	28,48	97,70	53
Albumin5	19,74	53,65	26
PPSB	27,70	81,01	31
AT3 (Import)	37,83	80,71	161
Coagulation Factors	34,96	97,04	28
GCSF	12,29	24,46	4
Insulin	1,31	4,91	208
Adrenaline Perfusor	1,54	14,60	69
Noradrenaline Perfusor	1,57	12,54	362
Dopamine Perfusor	1,31	10,63	264
Dobutamine Perfusor	1,38	9,47	273
Muscle Relaxants	19,80	68,85	174
Sedative Narcotic	1,68	9,50	396
O2AV	2,01	16,59	287
Peak	2,78	11,76	213
IE	1,30	7,748	245
Ventilation	1,07	2,90	399
Ventilation Rate	1,93	6,88	320
FiO2	1,23	5,63	389
PEEP	1,25	5,09	339
Haemofiltration	1,90	8,88	49
Dialysis	5,36	21,66	33
Organ Transplantation	36,70	0	1
Catheter	28,62	70,61	387
Parenteral Nutrition	2,43	7,01	402
Enteral Nutrition	3,85	12,34	273
SOFA	24,35	6,81	410
APACHE2	24,35	6,81	410
SAPS2	24,35	6,81	410
MODS	24,35	6,81	410
SIRS	23,82	10,67	411
HCL Solution	5,40	17,51	4
Albumin 20	32,10	97,24	99
Adrenaline Single	23,19	53,54	33
Noradrenaline Single	18,69	45,27	33
Dopamine Single	2,33	0,57	2
Dobutamine Single	1,69	2,52	8