in: A.Nierhaus, K.G.Kreymann (Eds.): Sepsis, SIRS, Immune Response – Concepts, Diagnostics and Therapy, Pabst Science Publ., Lengerich 2005, pp. 119-126

Predicting Death for Abdominal Septic Shock Patients- The Results of the MEDAN Project

R. Brause, E. Hanisch, J. Paetz, B. Arlt,

J.W.Goethe-University, Frankfurt a.M., Germany

1. Introduction

This contribution reviews the results of the MEDAN project – an analysis of a multi-center septic shock patient data collection. Since the description of sepsis by Schottmüller in 1914 [1], the amount on knowledge available on sepsis and its underlying pathophysiology has substantially increased. Epidemiologic examinations of abdominal septic shock patients show the potential for high risk posed by and the extensive therapy situation in the intensive care unit (ICU) [2]. Unfortunately, until now it has not been possible to significantly reduce the mortality rate of septic shock, which is as high as 50-60% worldwide, although PROWESS' results [3] are encouraging.

The heterogeneity of patients groups and the variations in therapy strategies is seen as one of the main problems for sepsis trials. Therefore, commonly available scoring systems are used for comparing critical ill patient groups. Moreover, one of the main objectives of scores is giving information with respect to outcome prediction. The task of the MEDAN project was the development of a septic shock diagnosis by self-learning systems, especially neural networks, and compare its performance to those of several established scores (SOFA, APACHE II, SAPS II, MODS). For this purpose, a group of 382 patients made up exclusively of abdominal septic shock patients for the first time in Germany was investigated by using scores and a multivariate neural network analysis. The classification results provided the basis for creating a reliable alarm system for abdominal septic shock patients.

2. Methods

For outcome prediction the data of 382 patients, who met the consensus criteria for septic shock [4],[5], were analyzed by using most of the commonly documented vital parameters and doses of medicine (metric variables). 187 of the 382 patients are deceased (48.9%). Data were collected in German hospitals from 1998 to 2001. All handwritten patient records were transferred to an electronic database. We used programmed range and plausibility checks of different kinds to detect all faulty data in the electronic database. Here, static values (e.g. lower and upper bounds) and dynamic development (e.g. time sequence behavior) were checked [6]. The complete data base is available at *www.medan.de/datenbank/download_ database.htm.*

For evaluation, we asked the question: Can we diagnose mortality very early or (trivially) only shortly before death or dismissal? How long before death or ICU dismissal can the outcome be predicted? Therefore, data from different periods of time are taken into account for the evaluation: the first three days of ICU stay (F3); the first three days after the septic shock occurrence (S3); all days of ICU stay (ALL); days 8,7 and 6, counted backwards from the last day of the ICU stay(D6-8), i.e. the last day of the ICU stay would be day 0; days 4, 3 and 2 counted from the last day of ICU stay (D2-4); and the last 5, 3, 2, 1 day(s) of ICU stay (L5, L3, L2, L1). All diagnosis results are characterized by their AUC value, the area under the ROC curve. Since the results on the admission day and the day after admission were almost random (AUC = 0.5) we used a minimum of three days (S3) for AUC calculation.

Interestingly, none of the traditional indicators could reach an acceptable level of diagnosis success. For instance, neither the doses of catecholamines nor the fact of enforced respiration is significant – because nearly all septic shock patients receive catecholamines and enforced respiration which seems to be a sign of septic shock and does not tell anything significant about the chances to survive. Also other single variables like base excess, lactate or O_2 saturation do not contribute to diagnosis. Even if they have a good diagnostic value like the central vein pressure or diastolic blood pressure, only in combination with other indicators the diagnosis becomes competitive. Therefore, the task became to find a small subset of the 140 variables with a good diagnostic power.

For this task, other data sets than those used by the scores are also taken into account, e.g. the 16 most frequently measured variables (*frequent16*), coagulation values (*coagulation*) like leukocytes, erythrocytes, haemoglobin, haematocrit, thrombocytes etc., heart system related variables (*heart*) like heart rate, systolic blood pressure, diastolic blood pressure, CVP etc., lung system related variables (*lungs*) like arterial pO_2, arterial pCO_2, base excess etc., breathing and catecholamines values (*bac*) like FiO_2, PEAK, respiratory frequency, adrenaline, noradrenaline, dopamine and dobutamine and the triple (*bpt*) of systolic and diastolic blood pressure and thrombocytes.

First, we computed the conventional, often used scores on the different data sets and evaluated their prognosis values on our data base and its subsets. The scores for comparison were:

a) SOFA (*Sepsis-Related Organ Failure Assessment*)[7],[8]. Ten variables are needed to calculate the score.

b) APACHE II (*Acute Physiological and Chronic Health Evaluation*) [9]. It uses a scale of 0 to 71 of whole-number values.

c) SAPS II (*Simplified Acute Physiology Score*) [10]: The SAPS II score is another ICU score using only 13 variables. Originally, SAPS was introduced as a simplified APACHE score.

d) MODS (*Multiple Organ Dysfunction Score*) [11]: The MODS score assesses organ states (respiratory, liver, renal, coagulation, heart, neurological) on a whole-number scale.

A score was calculated every time when the necessary variables were given. Generally we did not consider the Glasgow Coma Scale (GCS) [12] in the scores, because it was not always available in the data base.

Then, the neural network was trained on the data sets. Training was done with 50% of the samples and testing with the remaining 50%. In contrast to the predefined scores, the neural network algorithm [13] uses the known class information of the training data in its training process to obtain its diagnostic power. The outcome labels "survived" and "deceased" are used as class information in the training procedure of the neural network for its parameters, the weights. This kind of system adapts a non-linear classification to the data by adapting the position and width of rectangular basis functions in the input space. The classification is trained for optimal class discrimination and learns automatically to use the best subset of input variables to perform its task, avoiding the time consuming feature selection process. The result is similar to a nonlinear regression, but no regression model is needed a-priori. For implementation details see [14].

Finally, the samples of the test data sets are classified by the trained neural network. Data on training patients was not used for testing (disjoint patient sets). All experiments with one dataset were repeated twenty times for robust estimation of mean and standard deviation. As comparison criterion between the different performance results, the area beneath the ROC curve (AUC) is used.

Additionally, we computed the 95% confidence intervals (CI) for the AUC values of our neural network diagnosis by assuming that AUC values in one dataset calculated in repetitions of an experiment are normally distributed. Using explorative statistics (Q-Q-plots) this is a reasonable assumption. Therefore, the CI bounds can be obtained by linearly transforming the CI bound of a normal distribution using our measured variances and mean values, cf. [15], p.109.

3. Results

The results of the analysis of 382 patients are very similar to the intermediate results [16] obtained for only 138 patients. The three scores MODS, SAPS II and APACHE II perform differently when considering the last three days, time period L3 (Fig 1(a)), with APACHE II performing worst. Using the SOFA score (AUC = 0.90) results in a clearly better classification.

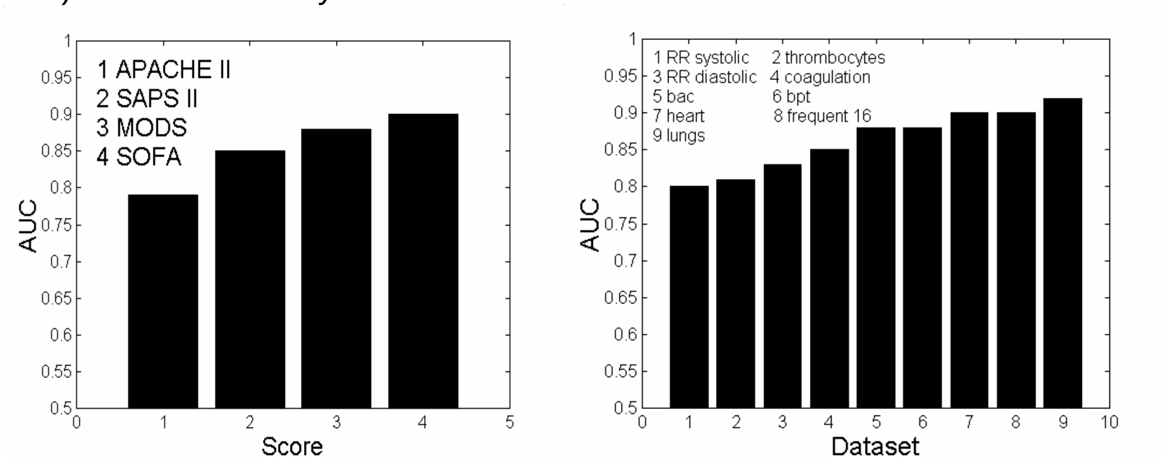

Figure 1: (a) Area beneath ROC curves (AUC) for MODS, SAPS II, APACHE II and SOFA (b) AUC values for different data sets for the last three days of ICU stay.

For the neural network, the AUC for different data sets (last 3 days of ICU stay L3) is shown in Fig. 1(b). Here, we obtain the same diagnostic power (AUC = 0.9) as the best score, using only the three variables of *bpt*. On the other hand, using roughly the same number of variables as the best score (SOFA), we obtain a better AUC for data set 8 or 9.

Now, for an early warning system ("alarm system") the question have to be answered: How early can a successful diagnosis be obtained? Considering the first three days (F3) the AUC for all scores and the network is very bad (near 0.5), and indicates the fact that no reliable diagnosis is possible on the first days of ICU stay. For different time periods the AUC values of the neural network diagnosis are plotted in Fig. 2(a), using the data set frequ16. The best classification results are achieved considering the last day L1. Since an outcome prognosis on the last day is not useful for building an alarm system we consider only the three-days prognosis horizon L3 with a high AUC of 0.9.

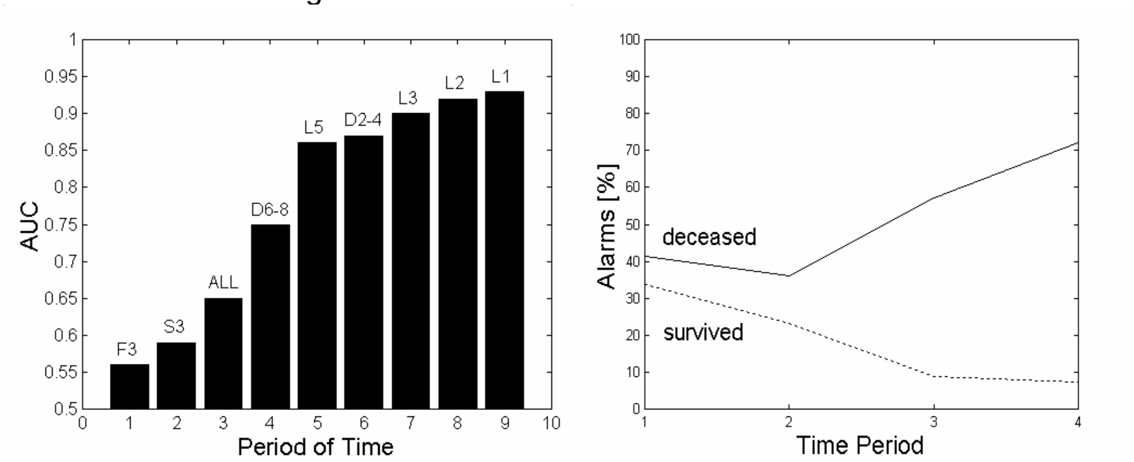

Figure 2: (a) Freqent16 data: AUC values for different periods of time of ICU stay are shown for the neural network diagnosis. (b) Alarm rate in percent for different time periods: the first three days(1); the first half of ICU stay (2); the second half of ICU stay (3); the last three days (4).

With the diagnostic results of the neural network we have created an alarm system [17], using 138 patients. Here, we present the results for the extended group of 382 patients. An alarm message is given whenever input for the neural network generates high output for class "deceased." In Fig. 2(b) we see the resulting alarm percentage for the first three days, for the first and second half of ICU stay and for the last three days, indicated separately for patients who either died or survived. In the time periods 1, 2, 3 and 4 the alarms decreased rapidly for surviving patients and increased for deceased patients. Only alarms (7%) stemming from the last three days can be interpreted as false alarms with respect to outcome prediction, because on the other days one cannot retrospectively examine if the alarms are due to critical or uncritical states which might occur independently. Alarms for survived patients might have not to be false; they might serve as indicators for critical periods of ICU stay.

4. Discussion

Most clinicians can recognize septic shock, but if you ask them, you get a hundred definitions [18], although consensus conferences should have resolved this issue [4],[5]. In this paper we use strictly the term „septic shock“, the term „severe sepsis“ is intentionally not applied, since in a former study we could demonstrate that „severe sepsis“ comprises almost identical patients with abdominal septic shock [19].

Different scoring systems have been developed, not only in order to document severity of illness, but also to estimate prognosis of critical ill patients. The best outcome predictor would be one that warns the physician on first day of ICU admission or when septic shock first appears (this is usually the second day of the patient's ICU stay according to our analysis). Our results demonstrate that none of the scoring systems achieves this goal. Only in the last three days of the ICU period, scores reach acceptable AUC values, whereby the SOFA score, based on ten variables, achieves the best AUC of all scores. Like the SOFA score, the data driven neural network approach performs similarly, using only three variables (*bpt*). For clinical practice, the good performance of the neural network can be obtained by a specially designed score, the MEDAN RRT score [16] of the three variables, see also *http://medan.de/scores/ scores.htm.*

Although scores and neural network under investigation provide relevant outcome prediction information only in the last three days of the ICU stay of patients (i.e. without clinical relevance) we found that scores are difficult to use for individual patients: a score value does not indicate death or survival with a high confidence resulting in long CIs. The neural network results on the non-score datasets are more reliable since CI length is usually shorter. The SOFA score has the lowest interval length (0.13) of all the scores. Therefore, it is the best score for abdominal septic shock patients from this point of view. For example, SOFA's CI length is 0.13, *bpt's* CI length is only 0.09. Considering all datasets (e.g. *lungs*, *heart*, *bpt*, *frequent16*), the results show the superiority of neural networks compared with scores when considering the confidence of a classification of individual patients.

The resulting alarm system based on our analyses produces reliable alarms: in the last three days of the ICU stay there were ten times more alarms for deceased patients then for survivors. The alarm system that was trained with data of the last three days represents the patient conditions that lead to death or survival with a high probability. Although the alarm system was trained with data of the last three days, it can be used as an online bedside alarm system. Right from the start of the patients' ICU stay physicians are warned when patients reach the same critical condition as deceased patients had within the last three days. If the patient is critical on his/her first day of ICU stay, the alarm system warns the physician, whether the patient will likely survive or die in the following days. If peripety happens later on, the alarm system will warn the physician at the right time.

In April 2002 a prospective randomized multicenter study was initiated to check the clinical usefulness of the web-based alarm system (see study protocol at *www.medan.de*).

Acknowledgements

This work was partially supported by the "Deutsche Forschungsgemeinschaft" DFG (German Research Foundation, project MEDAN: HA1456/7-1,2). The authors would like to thank all the participants of the MEDAN working group: all medical students involved in the project, the participating hospitals and all data typists. We thank Prof. Dr. A. Encke for supporting the MEDAN working group.

References

[1] Schottmüller H : *Wesen und Behandlung der Sepsis*. Inn. Med. 1914;31:257–280

[2] Hanisch E, Encke A : *Intensive Care Management in Abdominal Surgical Patients with Septic Complications*. In Faist E, ed., Immunological screening and immunotherapy in critically ill patients with abdominal infections, Berlin, Springer-Verlag, 2001;71–138

[3] Bernard GR, Vincent J-L, Laterre PF, et al. : *Efficacy and Safety of Recombinant Human Activated Protein C for Severe Sepsis*. N. Engl. J. Med. 2001;344:699–709

[4] Bone RC, Balk RA, FB Cerra, et al. : *American College Of Chest Physicians/Society of Critical Care Medicine Consensus Conference: Definitions for Sepsis and Organ Failure and Guidelines for the Use of Innovative Therapies in Sepsis*. Crit. Care Med. 1992;20:864–875

[5] Levy MM, Fink MP, Marshall JC, et al. : *2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference*. Crit Care Med, 2003;31 :1250-1256

[6] Paetz J, Arlt B, Erz K, Holzer K, Brause R, Hanisch E : *Data Quality Aspects of a Database for Abdominal Septic Shock Patients*, Journal of Computer Methods and Programs in Biomedicine, 2004, Vol 75(1), 23-30

[7] Vincent JL, Moreno R, Takala J, et al. : *The SOFA (Sepsis-Related Organ Failure Assessment) Score to Describe Organ Dysfunction/Failure*. Intensive Care Med. 1996;22:707–710

[8] Vincent JL, de Mendonca A, Cantraine F, et al. : *Use of the SOFA Score to Assess the Incidence of Organ Dysfunction/Failure in Intensive Care Units: Results of A Multicenter, prospective Study*. Crit. Care Med. 1998;26(11):1793–1800

[9] Knaus WA, Draper EA, Wagner DP, et al. : *APACHE II: A Severity of Disease Classification System*. Crit. Care Med. 1985;13(10):818–829

[10] Le Gall Jr, Lemeshow S, Saulnier F : *A New Simplified Acute Physiology Score (SAPS II) Based on a European / North American Multicenter Study*. JAMA 1993;270:2957–2963

[11] Marshall JC, Cook DJ, Christou NV, et al. : *Multiple Organ Dysfunction Score: A reliable Descriptor of a Complex Clinical Outcome*. Crit. Care Med. 1995;23(10): 1638–1652

[12] Jennett B, Teasdale G : *Assessment of Coma and Impaired Consciousness: A Practical Scal*e. Lancet 1974;1:81–84

[13] Paetz J : *Metric Rule Generation with Septic Shock Patient Data*. Proc. of the 1st IEEE Int. Conf. on Data Mining, 2001;637–638

[14] Brause R, Hamker F, Paetz J : *Septic Shock Diagnosis by Neural Networks and Rule Based Systems*. In Schmitt M, et al., eds, Computational intelligence processing in medical diagnosis. Heidelberg, Physica, 2002;323–356

[15] Hartung J : *Statistik*. Munich, Oldenbourg, 1998, 11th ed

[16] Brause R, Hanisch E, Paetz J, Arlt B : *The MEDAN-Project: Results and Their Medical Meaning,* 3rd Int. Symp. „Sepsis, SIRS, Immune Response – Concepts, Diagnostics and Therapy",A. Nierhaus, J.Schulte am Esch (Eds), PABST Science Publishers, Lengerich (Germany) 2003

[17] Paetz J, Arlt B : *A Neuro-Fuzzy Based Alarm System for Septic Shock Patients with a Comparison to Medical Scores*. Proc. of the 3rd Int. Symp. on Medical Data Analysis, 2002;42-52

[18] Rowe PR, Feature : *Septic Shock – Finding the Way Through the Maze*. Lancet 354, 9195, 1999.

[19] Wade S, Büssow M, Hanisch E : *Epidemiologie von SIRS, Sepsis und septischem Schock bei chirurgischen Intensivpatienten*, Chirurg 69, 1998; 648-655